Introduction to Bioinformatics Course

December 2002

Dr Andrew Lloyd (Ph.D.) and David Lynn (M.Sc.)

Supported by

The Conway Institute of Biomolecular and Biomedical Research,
University College Dublin.

&

The Dublin Molecular Medicine Centre (DMMC).

http://ercbinfo1.ucd.ie/course/bioinformatics2002.html

or

http://www.binf.org/course/bioinformatics2002.html

Table of Contents

Introduction to bioinformatics 3

Database formats and structure 6

Sequence formats & Accession Numbers 12

Day 1: Interrogating sequence databases: SRS Sequence Retrieval System at the EBI to find sequences by their annotation and Entrez.....15

Day 1: Nucleic Acid sequence analysis (Dec 2nd 2002).....22
- Translating DNA 23
- Reverse complement 23
- DNA analysis 24
- Primer design.....25
- Gene prediction 28
- Alternative splicing 30
- Promoter characterization.....34

Day 2: Protein Sequence Analysis (Dec 3rd 2002).....39
- Physico-chemical properties.....40
- Cellular localization 42
- Signal peptides 44
- Transmembrane domains.....47
- Post-translational modifications 50
- Motifs and domains 52
- Secondary structure prediction 53

Day 3: Accessing Complete Genomes (Dec 4th 2002) 56
- UCSC Genome Bioinformatics – The Golden Path 57
- Ensembl at EBI.....62
- NCBI Genomic Biology 64
- OMIM (ON-line Mendelian Inheritance in Man) 74

Day 4: Alignments and homology Searching (Dec 5th 2002) 75
- 2 sequence alignments.....75
- Similarity (homology) searching.....77
- Multiple sequence alignment.....88

Day 5: Phylogenetic trees Dec 6th 2002.....94
- Tree calculation methods: Distance matrix (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML). 95
- A PHYLIP protocol.....98
- Bootstrapping 100
- WebPhylip and PISE 105

Further readings in bioinformatics 114

Acknowledgements:
This course and manual grew naturally from The ABC Bioinformatics Course, an earlier project based on GCG and the WWW, to which Aoife McLysaght (formerly of TCD, currently at UC Irvine) was a major contributor. That in turn owes a debt of gratitude to the ABCT tutorial designed by Rodrigo Lopez when he was the Norwegian EMBnet node. This course would never have got off the ground without the encouragement of Cliona O'Farrelly, our boss and Research Director at the Education and Research Centre (ERC) at St. Vincent's Hospital. Any suggestions for improvement, notification of typos and the like should be sent on to david.lynn@ucd.ie or atlloyd@tcd.ie.

Introduction to Bioinformatics

This course is designed to impress upon you that computers and the Internet can not only make your work as a biologist easier and more productive but also enable you to answer questions that would be impossible without computational help. Thus there are some computational analyses that you could conceivably do on the back of an envelope or with a pocket calculator and there are others so computationally demanding that you would not attempt them without electronic help. An example of the first would be to scan the following DNA sequence for ecoRI restriction endonuclease sites (GAATTC):

```
>Adhr D.melanogaster
ATGTTCGATTTGACGGGCAAGCATGTCTGCTATGTGGCGGATTGCGGAGGGAGACCAGC
AAGGTTCTCATGACCAAGAATATAGCGAAACTGGCCATTCGGAAAATCCCCAGGCCATC
GCTCAGTTGCAGTCGATAAAGCCGAGTACTTCTGGACCTACGACGTGACCATGGCAAGA
ATTCATATGAAGAAGTACTGATGGTCCAAATGGACTACATCGATGTCCTGATCAATGGT
GCTACGCTGATAACATTGATGCCACCATCAATACAAATCTAACGGGAATGATGAACACG
TGTTACCCTATATGGACAGAAAAATAGGAGGAATTCGTGGGCTTATTGTTCGGTCATTG
GATTGGACCCTTCGCCGGTTTTCTGCGCATATAGTGCAGTGTAATTGGATTTACCAGAA
GTCTAGCGGACCCTCTTTACTATTCCCAGCTGTGATGGCGGTTTGTTGTGGTCCTACAA
GGGTCTTTGTGGACCGGGGTTTTTAGAATACGGACAATCCTTTGCCGATCGCCTGCGGC
GAGCGCCCCATCGGTTTGTGGTCAGAATATTGTCAATGCCATCGAGAGATCGGAGAATG
GATTGCGGATAAGGGTGGACTCGAGTTGGTCAAATTGCATTGGTACTCGACCAGTTCGT
GCACTATATGCAGAGCAATGATGAAGAGGATCAAGAT
```

(This sequence is written in Fasta format - see below for sequence formats.)

A computer could do it quicker, but it is still trivial to do it by eye. Especially as one of the sites has been picked out in bold. Can you find the other(s)? Sequence analyses impossible without a computer include, but are not limited to, most operations that involve the sequence databases. The DNA databases (Genbank EMBL DDBJ) are curated by three different groups in Bethesda, MD, Hinxton, UK and Mishima, JP but, because they exchange information on a daily basis, should be effectively the same in content. The DNA databases are doubling in size about every year; they are currently (07/11/02):

bases in size. So finding all of the ecoRI sites in the whole of a printed copy of the human genome (3Gbases) would take more than a few minutes.

This course will introduce you to some of the more commonly used bioinformatics tools, tell you how to use them and, more importantly, how to use them "correctly"

or at least more effectively. Most of the analysis will be carried out on the World Wide Web (WWW). This is partly because it is available to all comers without requiring direct access to the necessary computers, which serve as database and software repositories. But it is also partly because a well-designed Web site can be particularly user-friendly and intuitive in its operations.

There are likely to be network related problems trying to make 25 simultaneous connections over the Internet to the same site. We have scheduled the course for when the Internet is at its fastest. Try doing the course exercises late in the evening, early in the morning (best for speed!) or at weekends.

This 5 * half day module in bioinformatics is designed to give you a flavour of what analytical and informative tools are available on the World Wide Web.

Bioinformatics

Bioinformatics has been described as the storage, retrieval and analysis of biological sequence information. In this short course we will be taking a broader definition: how computers can maximise the biological information available to you. This will touch on determining the 3-D structure of bio-molecules and trying to relate this to their function as well as accessing the relevant literature. I hope that, by the end of the course, everyone will be adopting a more explicitly evolutionary understanding of 'their' molecule. The formal course practicals will be carried out entirely on the World Wide Web using Netscape or the other Web-browser.

You should note that several important types of bioinformatic analysis are not freely accessible on the Web, but are available on various password controlled computers. In particular, types of analysis that require large amounts of computational power/time are best carried out off the web. Analyses of many genes are also often better done in an environment where a computer program does the pointing and clicking for you. For the record, the GCG package is a suite of programs which carry out almost all the analyses that a molecular biologist might want to do with/on DNA or protein sequences (secondary structure prediction, two sequence alignment, conceptual translation of DNA, restriction site analysis, primer design, as well as homology searching, multiple sequence alignment etc.). For phylogenetic inference and tree drawing, the PHYLIP package (versions available for PCs, Macs and Unix) will answer most needs. Both of these software packages and a variety of other

sequence analysis packages are available at the Irish National Centre for BioInformatics (INCBI) contact Kevin Byrne (kbyrne@maths.tcd.ie) UCD also has a GCG site licence. GCG and PHYLIP are "packages" because they are internally consistent: if you have run one GCG program you can run any other.

The web, by contrast, is a total mess: the same program is implemented with different defaults at different sites; it is not often clear what those defaults/options/parameters are; the results are not easily transferred to a different program. It is free, but there is a cost! You are advised to validate any analysis against the results yielded by other sites.

Databases:

Databases are of course the core resource for bioinformatics. There is plenty of software for analysing one or a few sequences, but many of the computationally interesting and biologically informative programs access databases of information. Frequently used classes are the biological sequence databases. These include:

DNA databases:

- EMBL (European Mol Biol Lab)
- GenBank
- DDBJ (DNA DB of Japan)

The three DNA databases exchange their data on a daily basis and so should be identical as to content. They are, however, rather different in format:
Each of the database cited above consists of a (very large number) of entries, each consisting of a single sequence preceded by a quantity of 'annotation' that puts the sequence in its biological, functional and historical context. Without the annotation, GenBank would be a meaningless string of 22 billion As Ts Cs and Gs. Compare and contrast the two extracts from a) EMBL and b) Genbank (DDBJ has the same look-and-feel as Genbank):

a) EMBL

```
ID   ECRECA     standard; DNA; PRO; 1391 BP.
AC   V00328; J01672;
DT   09-JUN-1982 (Rel. 01, Created)
DT   12-SEP-1993 (Rel. 36, Last updated, Version 4)
DE   E. coli recA gene.
KW   .
OS   Escherichia coli
OC   Bacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
OC   Escherichia.
RN   [1]
RP   1-1374
RX   MEDLINE; 80234673.
RA   Sancar A., Stachelek C., Konigsberg W., Rupp W.D.;
RT   "Sequences of the recA gene and protein";
RL   Proc. Natl. Acad. Sci. U.S.A. 77:2611-2615(1980).
```

b) GenBank

```
LOCUS       ECRECA       1391 bp    DNA             BCT       12-SEP-1993
DEFINITION  E. coli recA gene.
ACCESSION   V00328 J01672
KEYWORDS    .
SOURCE      Escherichia coli.
  ORGANISM  Escherichia coli
            Eubacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
            Escherichia.
REFERENCE   1  (bases 1 to 1374)
  AUTHORS   Sancar,A., Stachelek,C., Konigsberg,W. and Rupp,W.D.
  TITLE     Sequences of the recA gene and protein
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 77 (5), 2611-2615 (1980)
```

You can see that these two are obviously talking about the same sequence from E.coli, but the information is encoded in a rather different way. This makes no difference to us reading the text, but causes problems when writing a program to interrogate a database. What do you think the EMBL codes OC and RT stand for?

Each database entry has a name, called ID or LOCUS, which tries to be mnemonic and marginally informative. More importantly each has an accession number which is arbitrary but which remains attached to the sequence for the rest of time. The organism might become reclassified, the gene may get renamed and the ID is thus subject to change, but by noting the accession number you should always be able to identify and retrieve the sequence. Note also that the original publication is cited. Usually there will be other papers documenting functional analysis, mutations, allelic variations, 3-D structure and so on.

Further down in the entry is annotation about the sequence itself, so that the sequence is parsed into meaningful bits called a features table:

a) EMBL

```
FT   source          1. .1391
FT                   /organism="Escherichia coli"
FT                   /db_xref="taxon:562"
FT   mRNA            191. .>1391
FT                   /note="messenger RNA"
FT   RBS             229. .233
FT                   /note="ribosomal binding site"
FT   CDS             239. .1300
FT                   /db_xref="SWISS-PROT:P03017"
FT                   /transl_table=11
FT                   /gene="recA"
FT                   /product="recA gene product"
FT                   /protein_id="CAA23618.1"
FT
FT   mutation        353. .353
FT                   /note="g to a in recA441 (E to K)"
FT   mutation        720. .720
FT                   /note="g to a in recA1 (G to D)"
```

b) GenBank

```
FEATURES             Location/Qualifiers
     source          1..1391
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     mRNA            191..>1391
                     /note="messenger RNA"
     RBS             229..233
                     /note="ribosomal binding site"
     gene            239..1300
                     /gene="recA"
     CDS             239..1300
                     /gene="recA"
                     /codon_start=1
                     /transl_table=11
                     /product="recA gene product"
                     /db_xref="SWISS-PROT:P03017"
     mutation        353
                     /gene="recA"
                     /note="g to a in recA441 (E to K)"
     mutation        720
                     /gene="recA"
                     /note="g to a in recA1 (G to D)"
```

Again you can see that the information exchange between Genbank and EMBL includes all significant portions of the annotation. Such useful signals and data as the open reading frame (CDS for CoDing Sequence), the ribosome binding site, intron boundaries, signal peptides, variants/mutations may be recorded.

Protein databases:

- SwissProt
- PIR (Protein Information Resource)
- GenPept

a) Swissprot

```
ID   RECA_ECOLI     STANDARD;      PRT;   352 AA.
AC   P03017; P26347; P78213;
DT   21-JUL-1986 (REL. 01, CREATED)
```

```
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   15-DEC-1998 (REL. 37, LAST ANNOTATION UPDATE)
DE   RECA PROTEIN.
GN   RECA OR LEXB OR UMUB OR RECH OR RNMB OR TIF OR ZAB.
OS   ESCHERICHIA COLI, AND SHIGELLA FLEXNERI.
OC   BACTERIA; PROTEOBACTERIA; GAMMA SUBDIVISION; ENTEROBACTERIACEAE;
OC   ESCHERICHIA.
...
...
CC   -!- FUNCTION: RECA PROTEIN CAN CATALYZE THE HYDROLYSIS OF ATP IN THE
CC       PRESENCE OF SINGLE-STRANDED DNA, THE ATP-DEPENDENT UPTAKE OF
CC       SINGLE-STRANDED DNA BY DUPLEX DNA, AND THE ATP-DEPENDENT
CC       HYBRIDIZATION OF HOMOLOGOUS SINGLE-STRANDED DNAS. IT INTERACTS
CC       WITH LEXA CAUSING ITS ACTIVATION AND LEADING TO ITS AUTOCATALYTIC
CC       CLEAVAGE.
CC   -!- INDUCTION: IN RESPONSE TO LOW TEMPERATURE. SENSITIVE TO
CC       TEMPERATURE THROUGH CHANGES IN THE LINKING NUMBER OF THE DNA.
CC   -!- DATABASE: NAME=E.coli recA Web page;
CC       WWW="http://monera.ncl.ac.uk:80/protein/final/reca.htm".
KW   DNA DAMAGE; DNA RECOMBINATION; SOS RESPONSE; ATP-BINDING; DNA-
BINDING;
KW   3D-STRUCTURE.
FT   INIT_MET      0      0
FT   NP_BIND      66     73       ATP.
FT   CONFLICT    112    112       D -> E (IN REF. 5).
FT   TURN          4      4
FT   HELIX         5     21
FT   HELIX        23     25
FT   TURN         29     30
```

etc etc

b) PIR

```
>P1;RQECA
recA protein - Escherichia coli
C;Species: Escherichia coli
C;Date: 31-Jul-1980 #sequence_revision 14-Nov-1997 #text_change 14-Nov-1997
C;Accession: G65049; A93847; A93846; S11931; S63525; S63979; A03548
...
C;Comment: The recA protein plays an essential role in homologous
recombination, in induction of the SOS response, and in initiation of
stable DNA replication.
C;Genetics:
A;Gene: recA
A;Map position: 58 min
C;Superfamily: recA protein
C;Keywords: ATP; DNA binding; DNA recombination; DNA repair; P-loop; SOS
response
F;67-75/Region: nucleotide-binding motif A (P-loop)
F;141-145/Region: nucleotide-binding motif B
F;73/Binding site: ATP (Lys) #status predicted
```

Note that these two entries refer to the same gene from *E.coli* despite differences in the way the data is encoded. However, in contrast to the difference between EMBL and Genbank, the quality of the annotation is quite different. The 3-D structure of this gene has been worked out and this information is reflected in the SwissProt entry as the position of every alpha-helix and beta-sheet is noted. In general, the quality of the annotation and the minimization of internal redundancy makes SwissProt the preferred database to use. However, note that PIR records the Genetic Map position of the gene; so it is probably good to scrutinize both databases to abstract maximal information.

SwissProt also gives added value by incorporating a large number of DR (database reference) tags, pointing to equivalent information in other databases.

a) SwissProt:

```
DR   EMBL; V00328; G42673; -.
DR   EMBL; X55553; -; NOT_ANNOTATED_CDS.
DR   EMBL; AE000354; G1789051; -.
DR   EMBL; D90892; G1800085; -.
DR   PIR; A03548; RQECA.
DR   PIR; S11931; S11931.
DR   PDB; 1REA; 31-OCT-93.
DR   PDB; 2REB; 31-OCT-93.
DR   PDB; 2REC; 01-APR-97.
DR   PDB; 1AA3; 23-JUL-97.
DR   SWISS-2DPAGE; P03017; COLI.
DR   ECO2DBASE; C039.3; 6TH EDITION.
DR   ECOGENE; EG10823; RECA.
DR   PROSITE; PS00321; RECA; 1.
DR   PFAM; PF00154; recA; 1.
```

When these are used as hypertext links they can enable a WWW browser to locate an extraordinary depth of detail about a given entry, 3-D structure (PDB), protein motifs (Prosite), families of related genes (Pfam), the DNA sequence (EMBL) and a couple of specialist E.coli added-value databases. SRS is one program that makes these hypertext links. The PIR cross-references are far fewer and less explicit; its reference to Genbank (GB:U00096) refers to the whole *E.coli* genome, whereas swissprot points specifically to the gene (DR EMBL; V00328)

b) PIR

```
...
A;Cross-references: GB:AE000354; GB:U00096; NID:g2367149; PID:g1789051;
UWGP:b2699
```

All these databases are made up of entries, concatenated one after the other in plain readable text. As such they are far bigger than necessary if you are trying to analyze the sequence rather than interrogate or browse the annotation. For these purposes, special high-compressed databases can be constructed. Frequently these are not readable by humans because they have been optimized for speed reading computers.

One of the simplest compression protocols is called Fasta format in which the annotation is edited down to a single title line followed by the sequence. The sequence at the top of the chapter is in Fasta format. All protein databases use the one-letter amino acid code, can you think why this might be?

Sequence Related Databases

Not all biologically relevant DBs consist of sequences and annotation. There are databases of journal abstracts, taxonomy, 3-D structures, mutations and metabolic pathways. Some of the most useful of these are databases which specialise in particular entities that can be found dispersed in the "whole sequence" databases.

You notice one of the cross-references for the SwissProt entry is:

DR PROSITE; PS00321; RECA; 1.

Prosite is a database of protein motifs. PS00321 is a family of proteins that all have the motif:

PA A-L-K-F-[FY]-[STA]-[STAD]-[VM]-R

and are all believed to bind DNA, hydrolyze ATP and act as a recombinase. One of the members of this family is the recA gene in E.coli which gives its name to PS00321. In the pattern above, the residues within [square brackets] are alternatives. Convince yourself that ALKFFAAVR could belong to the family but ALKFAAAVR could not. There are more than 1000 other families classified in a similar way. Finding a Prosite link in a Swissprot gene is a great help in finding other proteins related by structure and/or function.

Interpro - http://www.ebi.ac.uk/interpro/

You should also be aware of the Interpro project which incorporates and sorts data from a diversity of protein motif and domain databases into one searchable meta-database.

Sequence formats

As we have seen comparing database entries above, there are dozens of different ways in which you can store or represent the same fundamental information. Databases are often compiled in, highly conventionalised, readable English text. Computers, being not so bright, will have difficulty reading and interpreting the information unless the conventions are quite rigidly obeyed. There are a very large number of ways you can write, store and transmit simple one-dimensional sequence files. A common sequence interchange program called 'readseq' recognises at least 22 different file formats. If a computer program does not recognise the format of an input sequence it may not work or, worse, misinterpret header lines as sequence data or otherwise mangle your analysis. Some commonly used file/sequence formats are shown below:

1) GCG (a software package):

```
TRANSLATE of: ecrgcg check: 4152 from:
1 to: 1062 generated symbols 1 to: 354.

ECRECA.RECA           1062

ecrgcg.pep  Length: 354  Oct 15, 1998  Type: P  Check: 9572  ..
       1  MAIDENKQKA LAAALGQIEK
      51  ALGAGGLPMG RIVEIYGPES
     101  DPIYARKLGV DIDNLLCSQP
     151  TPKAEIEGE*
```

2) Fasta (named for a widely used homology searching program) – single title line beginning >:

```
>ECRGCG TRANSLATE of: ecrgcg  1 to: 1062
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

3**) Staden** (named after Rodger Staden - early, but still extant, software writer) – same as raw sequence:

```
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

4) NBRF/PIR (named after the protein database):

```
>P1;ecrgcg.pep
ecrgcg.pep, 354 bases, 218 checksum.
 MAIDENKQKA LAAALGQIEK
 ALGAGGLPMG RIVEIYGPES
 TPKAEIEGE*
```

Accession numbers

The information above makes you aware of the diversity of ways in which something so simple as a one-dimensional sequence may be represented. Another source of confusion is the variety of identifying numbers attached to sequences and knowing to which database they refer. Accession numbers are used as unique and unchanging numbers. They are not mnemonic, although databases also have a less stable, more memorable nomenclature: HBB_HUMAN, HSHBB, HUMHBB 2HBB are all human beta globin IDs in various databases,

- **GenBank/EMBL** accession numbers: originally a letter followed by 5 digits (X32152, M22239). When the number of sequences exceeded 2,600,000 - 2 letters followed by 6 digits (AL234556, BF345788).
- **SwissProt**. Still one letter followed by 5 digits, letter is either O,P,Q. P23445.
- **PIR**: the 'other' protein database, one letter followed by 5 digits, but numbers confusable with EMBL/GenBank: B93303 is chimp haemoglobin in PIR but a random genomic clone fragment in EMBL.
- **GenPept**. Conceptual translations from DNA that have not yet been annotated well enough to get into SwissProt. three letters and five digits, e.g.: AAA12345.
- **Trembl** (Translated EMBL): P or Q followed by 5 letters/digits.
- **PDB** protein structure records: 1 digit and three letters 1HBA, 1TUP

More recently, an attempt has been made to reduce the redundancy in the databases (there were 180 copies of D. melanogaster alcohol dehydrogenase each with its own accession number). One result is RefSeq - NCBI's "reference sequence" database

RefSeq: Two letters, and underscore bar, and six digits,

mRNA records (NM_*)	NM_000492
genomic DNA contigs (NT_*)	NT_000347
curated/annotated Genomic regions (NG_*)	NG_000567
Protein sequence records (NP_*)	NP_000483

We will see how RefSeq is becoming the central resource for gene characterization, expression studies, and polymorphism discovery. Because of the high level of necessary curation, it is not anywhere close to being comprehensive even for those species (human, mouse, rat) that are included.

Accession numbers give the community a unique label to attach to a biological entity, so we all know we are talking about the same thing. Sequences in databases evolve as their real biological counterparts do. They need to be updated, corrected and merged and we need to know which version of the sequence entry is being referred to. GenBank has used gi numbers and, more recently, version numbers for this. Each small change made to a genbank record gets the next gi number e.g. gi6995995 and so is totally arbitrary. Version numbers are appended to the accession number after a dot – V00234.2, NM_000492.2.

The other programs to use in the course are many and varied. We have tried to put links to them all on the course website:

http://ercbinfo1.ucd.ie/course/bioinformatics2002.html

OR

http://www.binf.org/course/bioinformatics2002.html

A few overall points for the course:

- Take the opportunity to compare and contrast different methods of doing a particular analysis.
- By all means take the defaults but be aware that changing them will almost certainly get more or better information.
- The Web is free, and you get what you pay for, so use the web with care & caution.
- As with lab work it takes time to get the protocol working. Once you have one that works for you, write it down, bookmark and remember it. But note, the web changes rapidly and you cannot afford to use outmoded technology for long.

DAY 1: Interrogating (sequence) databases

SRS - http://srs.ebi.ac.uk/

The DNA databases are enormously rich information resources partly because they are so big, but it would make little sense if it consisted of a long list of As Ts Cs and Gs. At the moment there are more than 3 million individual entries in EMBL. An entry could be a fragment as short as 3 base pairs (e.g. M23994) or a large contig consisting of many genes, including complete eukaryotic chromosomes (e.g. X59720). The value of the database lies substantially in the quality of the annotation which puts the sequence in its biological context.

As a biologist you may need to be able to interrogate the Database to find particular sequences or a set of sequences matching given criteria, such as:

The sequence published in Cell 31: 375-382
All sequences from Aspergillus nidulans
Sequences submitted by Peter Arctander
Flagellin or fibrinogen sequences
The glutamine synthase gene from Haemophilus influenzae
The upstream control region of Bacillus subtilis Spo0A

SRS (Sequence Retrieval System) is a very powerful, WWW-based tool, developed by Thure Etzold at EMBL and subsequently managed by Lion Biosciences, for interrogating databases and abstracting information from them.

One of the neatest features of SRS is the fact that interrelated databases can be cross-referenced with WWW hypertext links. This means that you can discover the protein sequence, the cognate DNA sequence, a family of related proteins in other species, a Medline reference to read an abstract of the original publication, a 3-D structure - all with a few point-and-clicks with the mouse.

There are several SRS servers on the Web. We will be using
http://srs.ebi.ac.uk/

at the EBI in England because a) it has a large number of interlinked databases b) connectivity to the UK is good c) they are attempting to interconnect their SRS server with their clustalW server and blast server. If the SRS server at the EBI is slow you might try any of:

http://srs.hgmp.mrc.ac.uk/

http://srs.sanger.ac.uk/

http://www.cmbi.kun.nl/srs6/

The latter is in the Netherlands and Irish connections to continental Europe can be very slow. The other three servers (EBI, HGMP and Sanger) are all located within a few metres of each other on the Wellcome Trust Genome Campus at Hinxton in England.

The documentation for SRS is getting better. With experience and practice you will get to use as much of SRS's power as necessary to obtain the results you need. I will show below, as a worked example, a series of instructions to obtain the sequences of all the mammalian osteonectin proteins in SwissProt, and download them locally to carry out a multiple sequence alignment using, say, clustalW. It should also be possible to do the multiple alignment on the EBI clustalW server (see day 4).

Use your browser (Netscape?) to go to http://srs.ebi.ac.uk/ or one of the other SRS servers at the top of the Course page.

You should see a page showing a stylized lion's paw.

- **Click the Start toe**

This gives you what is called the TOP PAGE. This page allows you to choose the database(s) that you wish to search. The databases may be of various types, including:

Sequence: Swissprot, sptrembl, PIR (Protein) or EMBL, emblnew (DNA)

Sequence related: prosite, blocks, prints (protein motifs and alignments), rebase (restriction enzymes),

Protein3Dstructure: PDB, HSSP

For more information about the contents of the database click on the relevant blue underlined hypertext link - Swissprot say.

- **Click the box [_] to the left of swissprot, or SWALL**

(if you don't know what SWALL contains, then click on the hyperlinked **SWALL** itself).

- **Click on the QUERY tab at the top of the page**

This will move you to a Query Form Page that permits you to submit particular queries (such as have been suggested at the beginning of this chapter) to the databases. At the top of this page will be a note of which database(s) you have chosen to search and a block of four text-insert boxes which you can use to enter your question.

to the left you will see five things you can change:

1. **[Reset]** - which clears the screen
2. **[Submit Query]** - which you have to click when you have entered your question
3. **Append wildcard to words [_]** which means that "bact" will be interpreted as bact* and look for bacteria, bacteriophage, etc.
4. **combine searches with [AND]** - which enables you to apply other logical (boolean) operators
5. **Number of entries to display per page** (default is 30)

Your question can be entered into one of more of the text-insert boxes, thus:

- **Click [All text] change to [Description]** and **insert osteonectin in box**

Note: it does not have to be osteonectin it could be ubiquitin or haemoglobin or hemoglobin or actin & alpha. Separate keywords in the same box have to be linked by a logical (Boolean) operator such as

and: **&**

or: |

but not: **!**

- **Click next [All text] change to [Organism] and insert mammalia in box**
- **Click [Submit query]**

a new window appears with query "[swall-Description:osteonectin*] & [Swall-Organism:mammalia*]" found 8 entries towards the top. This is how SRS interprets what you have entered in the boxes and the numbers of "hits" found.

- **Click [SequenceSimple] change to [FastaSeqs]**
- **Click [View]**
- **Click [Save]**

Make sure use view is [FastaSeqs] and sequence format is [fasta]

- **Click [Save]**
- **Click Netscape's File ....Save As**
- **Click Format for Saved Document [source] change to [text]**
- **Change selection .../wgetz to .../osteo.pro and then Click [Okay]**

This should dump the concatenated fasta format protein sequences into a local file called osteo.pro. You can use this file as input for clustalw (say in week 4 of this course). There may be local security difficulties with downloading sequences onto a public terminal - check with your neighbours or your demonstrator.

Query manager: a powerful tool

A quick example will show how you can combine very complex queries to zero in on the sequence(s) you need.

Having selected your database(s) go to the Query Form Page and enter:

- **[Description] calmodulin**

you should get about 800 entries.

- **Click [QUERY] tab** at the top of the page to get a new page and enter:
- **[Organism] human** (or indeed Homo sapiens)

this will get you a large (~50,000) number of sequences.

- **Click [RESULTS] tab**

A new window should appear with the results for all the queries you have entered in the current SRS session. In the top box of this page **enter "Q1 & Q2"** (leave off the quotes!) **Note**: *Your mileage may vary here. Q1 and Q2 may refer to earlier queries in this SRS session (osteonectin?) so use good judgement*.

- **Click [Expression]** box to the left of the query-entry box.

You have just used a boolean logical expression to yield about 80 sequences which are a) human and b) have "calmodulin" in the SwissProt description. This shows

you how it can be unreliable to depend on the annotation to get homologous sequences. Nevertheless, the list should contain the SwissProt entry for CALM_HUMAN which is what you want.

Questions

1. Can you think of a better way to find other mammalian calmodulin genes ?
2. If you do a search in SwissProt for "calmodulin" using the [AllText] descriptor instead of [Description] you find many more entries, why do you think you get more entries under this search?
3. There are more entries in SwissProt under [Organism] dog than [Author] dog, but more for [Author] wolf than [Organism] wolf. Why do you think this is so ?
4. Searching [Organism] mouse in SwissProt yields some plant sequences: prove this by finding sequences matching [Organism] mouse & [Organism] viridiplantae. Why is this so? (Clue: append wildcard *).

You should be able to reveal the full SwissProt entry for any protein sequence. If you do this you will see several (? blue, underlined) hypertext links to related databases. Almost certainly at least one of these will be EMBL and one to Medline. Probably one will be the prosite motif database. If the 3-D structure is known, one link will be to PDB. Investigate these other databases to get as much relevant information as possible about your sequence.

It is this, interlinked databases, aspect of SRS which gives it a large part of its power. You can extend your search to include other sequences related in some particular (or peculiar!) way. The Prosite link allows you to find members of a protein family. The EMBL link allows you to find the introns and the intron splice junctions, not to mention the ribosome-binding site, the stop codon and the journal reference for the original sequence. The Medline link will give you an abstract etc. You will probably find that:

The PubMed server at http://www.ncbi.nlm.nih.gov/Entrez/ is a far better tool for browsing Medline that what is offered with SRS. Especially powerful is its facility for finding [Related entries]. We will be covering PubMed in week 4 of the course).

Additional questions:

1. Who submitted the serum amyloid A (SAA) gene sequence for *Canis familiaris*?
2. What prosite motif defines the recA family of prokaryotic proteins? Which Dublin-based phylogeneticists used mult-sequence alignment to define this motif?
3. What are the first and last 5 bases in the intron of the yeast actin gene with EMBL accession number V01288?
4. What is the map position of one of the human SAA genes (SwissProt: P02735)? What cross-reference database is most likely to have map position?
5. What mutation at what position causes phenylketonuria (PKU)? (hint: EMBL K03020) but then try SwissProt: P00439.
6. What bases define the ribosome binding site of the Bacteroides fragilis glnA gene? Perhaps start from the E.coli homolog SwissProt: P06711.
7. Why is the name Saarinen associated with life-threatening cardiac arrythmias? (Hint: not because of architectural flaws...try voltage gated potassium channels)
8. Are there more publicly available DNA sequences from Rodents or Prokaryotes? What about protein sequences?
9. Get a sample of mammalian introns. See what common features they have? Think how these common features might help splicing out the introns.

The terminals in Botany are configured so that you can display, more of less directly, 3-dimensional structures from SRS. Back home you will need to get a copy of the RasMol 3-D structure viewer and install it in such a way that your Netscape/IE will recognise it and connect suitable (3-D sequence) file to it.

To display a PDB entry of 3-D coordinates as a rotatable, colorable model you need to click on the **[save]** button. The change the "use mime type" choice-box to **chemical/x-pdb** and then click on the **[save]** box. This should fire up CHIME a WWW implementation of RasMol.) Your mileage may vary!

Entrez - http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is the US equivalent of SRS and is available from the NCBI webpage You will most likely be familiar with Entrez for interrogating Medline, but the same engine can be pointed at DNA and protein databases. It is handy if you are familiar with the Entrez system and you want a sequence whose name or accession number you already know. At the top of the Entrez page change the Search [__] choice box from PubMed to the appropriate sort of database – the available options are listed on the Entrez page. If you want the sequence alone – to paste into some analysis page – change the Display [__] choice box to FASTA then click on [Save] or [Display] depending on whether you want a permanent or transitory copy of you proteins. Entrez has a more complex syntax for less straight forward queries.

DAY 1: Nucleic Acid Sequence Analysis

TOPICS:

1. Translating DNA in 6 frames.
2. Reverse complement & other tools.
3. Calculating some properties of DNA/RNA sequences.
4. Primer design.
5. Gene prediction.
6. Alternative splicing.
7. Promoter characterisation.
8. Other resources.

1) Translating DNA in 6-frames:

Translate tool - http://www.expasy.ch/tools/dna.html

This tool allows the 6-frame translation of a nucleotide (DNA/RNA) sequence to a protein sequence in order to locate open reading frames in your sequence.

- Go to URL above.
- Paste your sequence in the box provided & click "TRANSLATE SEQUENCE".
- You can choose 3 options
    - Verbose – puts Met & Stop to highlight start & stop codons.
    - Compact – useful if you want to use output in other programs.
    - Includes nucleotide sequence – nucleotide sequence is above the translation.
- This returns a 6-frame translation of your sequence. You can then choose the correct frame.

See Appendix II for the genetic code

2) Reverse Complement & other tools:

There are many cases where you might want to obtain the reverse complement of a DNA sequence, for example the reverse complement is needed as a negative control when doing a DNA hybridisation experiment.

Search launcher at Baylor College –
http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html

This tool contains a number of different applications for nucleic acid sequence analysis: For each application you can click on the following [H] [O] [P] [E] = [H]:Help/description; [O]:full Options form; [P]:search Parameters; [E]:Example search. On all the Baylor pages (and everywhere else possible) it is important to investigate the options [O] to see a) what are the defaults and b) what options seem worth changing. The following programs are available:

Readseq:

Converts nucleic acid/protein sequences between any of 30 different formats. It is often appropriate to convert to FASTA format. A large number of input formats are permitted. See help for details [H].

RepeatMasker:

RepeatMasker is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes as well as for low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (replaced by Ns). On average, over 40% of a human genomic DNA sequence is masked by the program. This is important in primer design so that you do not design a primer that spans a region with repeats. It is also important before doing a homology search as repeats in your sequence may hit other repeats in the genome (although BLAST now does this for you).

Primer Selection -PCR primer selection (See primer design later).

WebCutter- restriction maps using enzymes w/ sites >= 6 bases.

6 Frame Translation - translates a nucleic acid sequence in 6 frames.

Reverse Complement - reverse complements a nucleic acid sequence.

Reverse Sequence - reverses sequence order.

Sequence Chopover - cut a large protein/DNA sequence into smaller ones with certain amounts of overlap.

HBR - Finds *E.coli* contamination in human sequences.

Exercise: Paste in your own sequence of interest or alternatively examine an example output for each application by clicking [E] beside each program. Pay particular attention to the options available: these will give you clues about standard practice.

3) Oligo Calculator - http://www.pitt.edu/~rsup/OligoCalc.html

Tool to calculate the **length**, **%GC content**, **Melting temperature (Tm)** the midpoint of the temperature range at which the nucleic acid strands separate, **Molecular weight**, & what an **OD** = 1 is in picoMolar of your input nucleic acid sequence.

Many of these parameters are useful in primer design (see next section) and in other areas of molecular biology.

- Go to URL above.
- Paste your sequence in the box provided & click "Calculate".

Example:
>gi|10834993|ref|NM_000641.1| Homo sapiens interleukin 11 (IL11), mRNA
Length = 2281
% GC content = 55
Tm = 87 °C
Molecular Weight = 704856 daltons (g/M)
OD of 1 = 41 picoMolar

4) Primer design - Originally written in Jan 2002 by Norma O'Donovan (Thanks!).

The recommended site (although there are several others available on the web) is

Information about GeneFisher:

http://bibiserv.techfak.uni-bielefeld.de/genefisher/help/wwwgfdoc.html

The submission form:

http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/gf_submit?mode=STARTUP&sample=dna

The input form is straightforward and well documented.

Xprimer (Primer Selection) - very easy to use but not as good as GeneFisher:

http://alces.med.umn.edu/rawprimer.html

Primer Design Tips:

- **Primer Length:** usually between 18 and 24 base pairs
- **% GC**: Optimum GC content is 45-55%
- **Annealing Temperature:** Should be between 55°C and 65°C and ideally the annealing temperature of the 2 primers should be similar. A quick equation (Wallace formula) for calculating the annealing temperature of the primer is:

2 x (no. of As + Ts) + 4 x (no. of Gs + Cs)

The lower of the 2 primer annealing temperatures is the highest temperature that can be used for annealing. (Usually when optimising PCR you would start with an annealing temp. a few degrees below the Tm of the primers).

- **G/C clamps:** The 3' end of the primer should be able to form G/C clamps, i.e. several consecutive G/C or C/G base pairs, between the 3' end of the primer and the template DNA.
- **Length of PCR product:** The optimum size is 100 – 500 base pairs for conventional PCR. Shorter products can be used for real time PCR or longer products can be amplified using special polymerases.

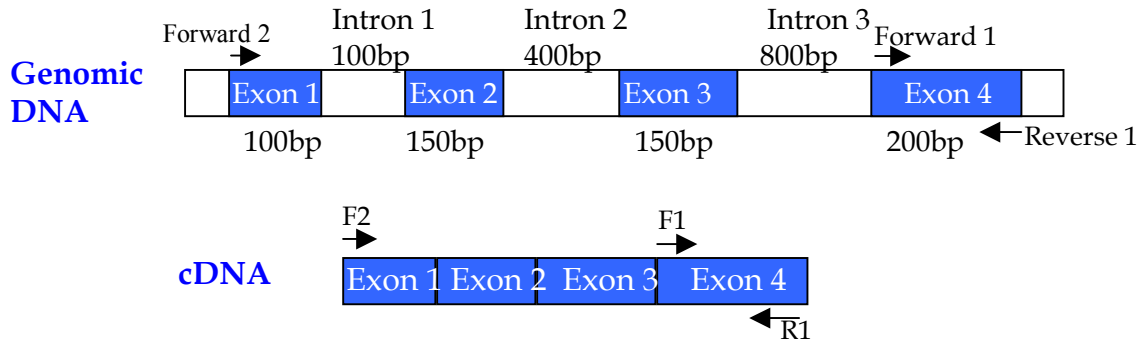
Things to avoid:

1. Complimentarity within a primer or between 2 primers (especially in the ends), used in the same reaction, as this may cause primer dimers.
2. Strings of a single nucleotide (more than 3).
3. Non-specific binding of primers to related sequences – check the specificity of the primers by doing a BLAST search of the database (non-redundant and genomic) with each of the primer sequences.

Primers for RT-PCR

The same rules as above apply but there are a few extra considerations. If you are doing RT-PCR with total RNA there may be genomic DNA contamination present in the RNA. (You can DNase treat to remove it or purify poly-A mRNA). If it is not removed you must ensure that your primers specifically amplify the cDNA (complementary to mRNA). Ideally the primers should not amplify the genomic DNA at all but if that is not possible the genomic product should be distinguishable from the cDNA product on a gel, based on size. Therefore the primers must span at least one intron in the genomic DNA. To identify the position of introns in the sequence align the mRNA sequence with the genomic sequence using a pairwise BLAST sequence alignment (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html). Alternatively, for human or mouse sequences, on the UCSC website (http://genome.ucsc.edu/) you can do a BLAT search with the mRNA which will identify the intron/exon structure of the gene.

Example:



If the forward and reverse primers are designed in exon 4 the PCR product obtained from the cDNA will be the same size as the genomic PCR product. If the forward primer is in exon 1 and the reverse primer is in exon 4 the cDNA product will be approx. 600bp whereas the PCR product from genomic DNA would be about 1900bp, which probably wouldn't be amplified in conventional PCR.

5) Gene Prediction

Gene prediction is an area under intensive research in bioinformatics and an entire course could be dedicated to it alone. Here we will introduce the GENSCAN program since it was one of the major programs used to predict genes in the human genome. This program should be useful in predicting genes in most vertebrate species, although caution should be used when dealing with other species especially prokaryotes where other programs are more suitable.
For links to other programs used in gene prediction take a look at:

The Institute for Genomic Research - http://www.tigr.org/software/
OR
The Deambulum Nucleic Acids Sequence Analysis page at Infobiogen -
http://www.infobiogen.fr/services/deambulum/english/prog2.html#STRUC

GENSCAN - http://genes.mit.edu/GENSCAN.html

Program designed to predict complete gene structures, including exons, introns, promoter and poly-adenylation signals, in genomic sequences.
For information about Genscan, click here link. It will tell more…

- Paste your sequence in the box provided & change the "print options:" to "Predicted CDS and peptides". Other defaults are OK.
- Click "Run GENSCAN".
- The Genscan prediction for the genomic region around human hepcidin is shown below. Genscan predicts that the initial exon begins at position 2566 and ends at 2477, then there is an intron, then there is the 2nd exon (355 – 296). The 3rd exon is at 206bp – 102bp. Genscan also predicts the termination signal and the polyA tail. The gene is on the complementary strand.
- The predicted coding sequence (CDS) & predicted peptide match exactly the known sequences for hepcidin.
- READ the output carefully then try your own sequence. It should be reasonably self-explanatory.

Example: chr19:42030157-42032793 genomic sequence of human hepcidin.

GENSCANW output for sequence 09:24:40

```
GENSCAN 1.0      Date run: 17-Dec-101      Time: 09:24:40
Sequence 09:24:40 : 2637 bp : 53.92% C+G : Isochore 3 (51 - 57 C+G%)
Parameter matrix: HumanIso.smat

Predicted genes/exons:
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.04 PlyA -     31     26    6                                1.05
 1.03 Term -    206    102  105  2  0  118   42    49 0.878   2.01
 1.02 Intr -    355    296   60  1  0  119   89   -16 0.646   1.02
 1.01 Init -   2566   2477   90  1  0   99   94   173 0.999  17.44

Click here to view a PDF image of the predicted gene(s)
Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

Predicted coding sequence(s):

>09:24:40|GENSCAN_predicted_peptide_1|84_aa
MALSSQIWAACLLLLLLLASLTSGSVFPQQTGQLAELQPQDRAGARASWMPMFQRRRRRD
THFPICIFCCGCCHRSKCGMCCKT

>09:24:40|GENSCAN_predicted_CDS_1|255_bp
atggcactgagctcccagatctgggccgcttgcctcctgctcctcctcctcctcgccagc
ctgaccagtggctctgttttcccacaacagacgggacaacttgcagagctgcaaccccag
gacagagctggagccagggccagctggatgcccatgttccagaggcgaaggaggcgagac
acccacttccccatctgcattttctgctgcggctgctgtcatcgatcaaagtgtgggatg
tgctgcaagacgtag
```

Explanation

Gn.Ex : gene number, exon number (for reference)
Type : Init = Initial exon (ATG to 5' splice site)
Intr = Internal exon (3' splice site to 5' splice site)
Term = Terminal exon (3' splice site to stop codon)
Sngl = Single-exon gene (ATG to stop)
Prom = Promoter (TATA box / initation site)
PlyA = poly-A signal (consensus: AATAAA)
S : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exon or signal (numbered on input strand)
End : end point of exon or signal (numbered on input strand)
Len : length of exon or signal (bp)
Fr : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph : net phase of exon (exon length modulo 3)
I/Ac : initiation signal or 3' splice site score (tenth bit units)
Do/T : 5' splice site or termination signal score (tenth bit units)
CodRg : coding region score (tenth bit units)
P : probability of exon (sum over all parses containing exon)
Tscr : exon score (depends on length, I/Ac, Do/T and CodRg scores)

Comments

The SCORE of a predicted feature (e.g., exon or splice site) is a log-odds measure of the quality of the feature based on local sequence properties. For example, a predicted 5' splice site with score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor (more than likely not a real donor site).

The PROBABILITY of a predicted exon is the estimated probability under GENSCAN's model of genomic sequence structure that the exon is correct. This probability depends in general on global as well as local sequence properties, e.g., it depends on how well the exon fits with neighboring exons. It has been shown that predicted exons with higher probabilities are more likely to be correct than those with lower probabilities.

6) Splice site prediction / Alternative splicing

Introduction to splicing:

Taken from http://www.bioinformatics.ucla.edu/HASDB/

The first requirement for proper splicing is some way to distinguish exons from introns. This is accomplished using certain base sequences as signals. These consensus base sequences, as they are known, allow the spliceosome (the cellular machinery that does the splicing) to identify the 5' and 3' ends of the intron. For example, in eukaryotes, the base sequence of an intron begins with 5' GU, and ends with 3' AG. [Figure] These sequences base pair with complementary spliceosomal RNA so that the pre-mRNA is aligned properly with the spliceosome. Each species has additional bases associated with these splice sites, but GU and AG are the only ones that are conserved across all eukaryotes. For example, the consensus sequence at the 5' splice site of vertebrate introns is AG***GU***AAGU (Stryer, 1995). Introns also have another important sequence signal called a branch site containing a tract of pyrimidine bases and a special adenine base, usually approximately 50 bases upstream from the 3' splice site. More information on the mechanism of splicing is available at the above website but will not be discussed in this course.

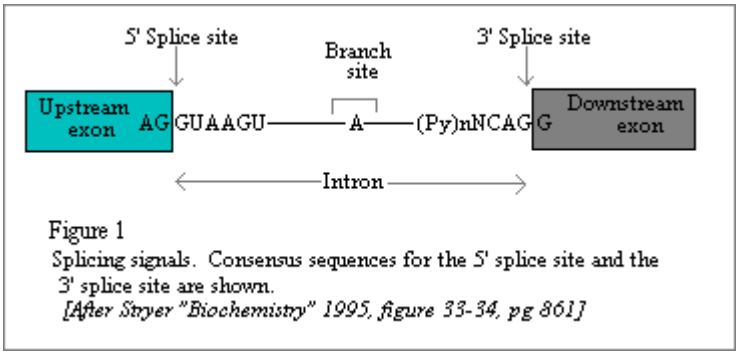


Figure 1
Splicing signals. Consensus sequences for the 5' splice site and the 3' splice site are shown.
[After Stryer "Biochemistry" 1995, figure 33-34, pg 861]

Alternative splicing:

The central dogma of molecular biology was that 1 gene = 1 protein, however more and more examples have been discovered where this is not the case and multiple possible mRNA transcripts can be produced from 1 gene and if translated these transcripts can code for very different proteins. This phenomenon is known as alternative splicing. There are 4 basic ways in which alternative splicing can occur:

1) Splice / Don't Splice

First, an intron can either be spliced out of the RNA (as in the simple model of RNA splicing), or it can be retained and included in the coding region of the RNA. This phenomenon is known as splice/ don't splice and the choice could have several different results. For example, if the intron includes an in-frame stop codon, then a splice variant that includes the intron may result in a shorter, non-functional protein. If the intron is spliced out, then the resultant mRNA would have an open reading frame which would be translated into the functional protein. In this case, the alternative splicing acts like an on/off switch. Another potential outcome of splice/ don't splice is simply that two functional mRNAs could be made, each with a unique base sequence. This would create two different proteins, each with a unique amino acid sequence, and possibly with different but related functions. In this case, the alternative splicing acts like a switch between producing mRNAs coding for two different proteins.

2) Competing 5' or 3' Splice Sites

A second mechanism for alternative splicing is the presence of competing 5' splice sites for one 3' site within one intron. Alternatively, there can be competing 3' splice sites for one 5' site within one intron. The competing site that is closest to the other end of the intron is called the proximal site, while the competing site that is farthest from the other end of the intron is called the distal splice site. The selection of each splice site would result in mRNAs that differed by the stretch of bases between the proximal and distal splice sites. Like the possible outcomes of splice/ don't splice, competing 5' or 3' sites could act like an on/ off switch, or this mechanism could act like a switch between the production of mRNAs coding for two different proteins.

3) Exon Skipping

A third mechanism for alternative splicing is called exon skipping. This occurs when an exon that would usually be included in the mature mRNA is spliced out with the neighboring introns, and is therefore skipped. There can also be multiple exon skipping in which more than one exon (with intervening introns) is skipped at once. This mechanism has the potential to produce many different mRNA's. For example, if a gene has 8 exons, one variant might include all of them, while another variant skips exon 7, and another variant skips exons 2 and 3, and yet another variant skips exons 4 and 5, etc... Hence, exon skipping has the potential to lead to many different mRNAs that could function as on/ off switches or as a switch between maturation of mRNAs for different proteins.

4) Mutually Exclusive Exons

A mechanism of alternative splicing related to exon skipping is called mutually exclusive exons. In this case, the mRNA would include either exon 1 or 2, not both. For example, if a gene has 4 exons, one splice variant might include exons 1, 2 and 4, while another splice variant might include exons 1, 3 and 4. Again, there is the potential for an on/off switch and for a switch between mRNAs for two proteins. It is important to note that more than one of these modes of splicing could happen at the same time. For example, it is possible that a gene could be alternatively spliced through both exon skipping and competing 5' splice sites at the same time. It is also important to note that research into alternative splicing is in the early stages, and that other modes of alternative splicing may be discovered in the future.

The Human Alternative Splicing Database at UCLA –

http://www.bioinformatics.ucla.edu/HASDB/

Used ESTs to locate alternative splices. Project has resulted in a publication of over six thousand alternatively spliced isoforms of human genes.

You can search the database using any of the following identifiers:

- **Gene Symbol:** search by a gene symbol (e.g. TCN1)
- **UniGene Sequence Identifier:** search by a UniGene sequence identifer (e.g. Hs#S3362)
- **UniGene Cluster Identifier:** search by a UniGene cluster identifier (e.g. Hs.2012)
- **Gene Title:** search by a gene title (e.g. transcobalamin I (vitamin B12 binding protein, R binder family))
- **GeneBank Sequence Identifier:** search by a GeneBank sequence identifier (e.g. J05068)

You can also search for tissue-specific alternative transcripts by clicking "Search By Tissue".

Example: HLA-G (gene symbol)

HLA-G is a nonclassical MHC 1 molecule that inhibits NK cell function. At least 7 variants have been characterized and these variants may have very different functions. Search HLA-G at HASDB to view the variants determined by this project.

NOTE: On day 3 we will see how the genome browser at UCSC can be used to help identify alternative splice variants.

7) Promoter Analysis & Recognition:

A promoter is a sequence that is used to initiate and regulate transcription of a gene. Most protein-coding genes in higher eukaryotes have polymerase II dependent promoters.

Features of pol II promoters:

- Combination of multiple individual regulatory elements.
- Most important elements are transcription factor binding sites.
- CAAT or TATA boxes are neither necessary nor sufficient for promoter function.
- In many cases, order and distances of elements are crucial for their function.
- Sequences between elements within a promoter are usually not conserved and of no known function.

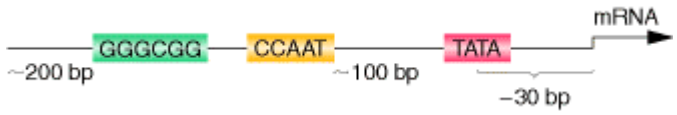


Figure 14-19: Taken from "Modern Genetic Analysis" (W.H. Freeman & Company).
The promoter region in higher eukaryotes. The TATA box is located approximately 30 base pairs from the mRNA start site. Usually, two or more promoter-proximal elements are found 100 and 200 bp upstream of the mRNA start site. The CCAAT box and the GC-rich box are shown here. Other upstream elements include the sequences GCCACACCC and ATGCAAAT.

Promoter identification

Polymerase II promoters are generally defined as the region of a few hundred base pairs located directly upstream of the site of initiation of transcription. (More distal regions and parts of the 5' UTR may also contain regulatory elements and may be part of the promoter). The exact length of a promoter can often only be defined experimentally. However, for an initial in silico analysis it may be sufficient (and also necessary) to restrict the region to about 300 to 1000 bp upstream of the transcription start site. Therefore, identification of the transcription start site directly leads to the location of the promoter of a gene. The transcription start site can be defined by mapping a 5' full-length mRNA/cDNA (including the complete 5' UTR) to the genomic sequence. The second possibility is to use PromoterInspector, a tool that is able to predict promoter regions in genomic sequences.

PromoterInspector –

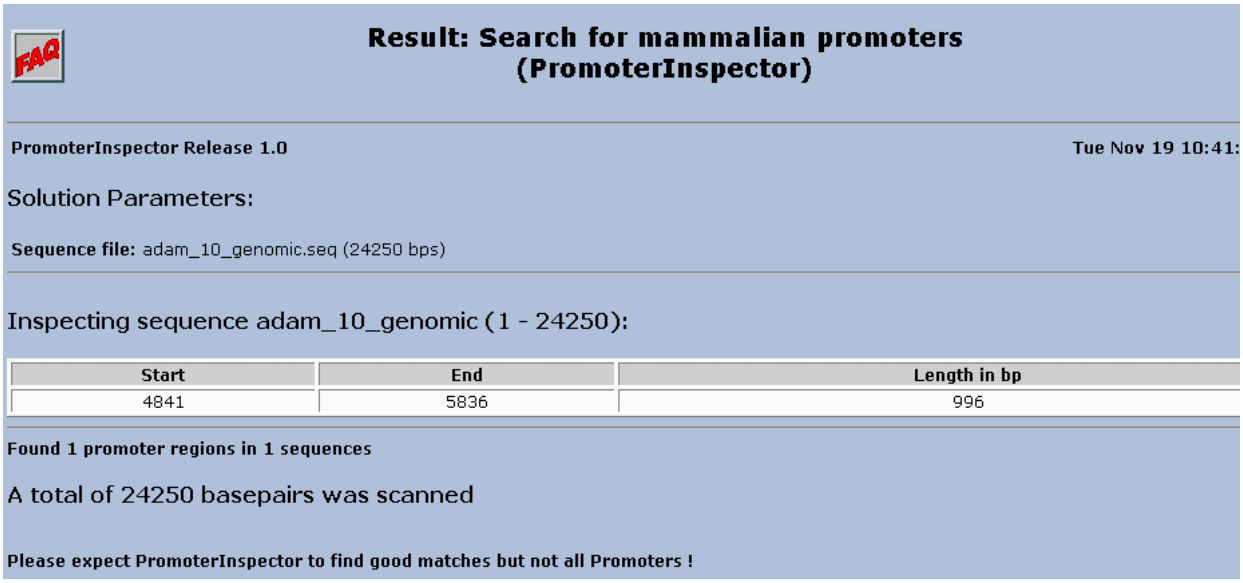
http://www.genomatix.de/software_services/software/PromoterInspector/PromoterInspector.html

PromoterInspector is a program that predicts eukaryotic pol II promoter regions with high specificity (~ 85%) in mammalian genomic sequences. PromoterInspector focuses on the genomic context of promoters rather than their exact location. The sensitivity of PromoterInspector is about 50% which means that the current version predicts about every second promoter in the genome. Therefore, your promoter may not be found. PromoterInspector predicts the approximate location of a promoter region and not the exact location of the Transcription Start Site (TSS). The predicted regions may contain the promoter or overlap with the promoter. The strand orientation of the predicted promoter region can only be derived from the location of the corresponding gene. PromoterInspector predicts promoter regions by identification of the conserved promoter context independently of the occurrence of specific elements like CCAAT or TATA boxes. To identify transcription factor binding sites in a promoter you can use MatInspector professional (see below).

- Go to the genomatix link above.
- Scroll down page & Click "Use PromoterInspector online".
- Before you can use this program you will have to register and obtain a user name and password. Do this by filling in the form & clicking "Register now!"
- Once you have obtained a user name and password by e-mail you can use PromoterInspector.
- There are two ways to supply a set of input sequences: Either enter your sequences directly into the form or, if your browser supports this option, a sequence file can be uploaded.
- In both cases, the input sequences must be in one of the supported formats.
- Please note that the web version of PromoterInspector accepts up to 100000 base pairs sequence input! To keep computing times reasonable the program should NOT be used more than 3 times per day per user!
- Supply your VALID email address & click "Start PromoterInspector".
- When the analysis is finished, an email with the URL of the results will be sent to this address. You can then point your browser to this address.

- The result will be available for 30 days on the server. After that period it will be deleted! Depending on the number and length of sequences in your input, the computation may take a while.
- PromoterInspector creates an output file that contains a list of promoter regions for every sequence. Start and end of the regions correspond to sense strand numbering. **Please note that predictions are not strand specific**.

Example: >chr15:56167697-56191947 (reverse complemented) genomic sequence around the human ADAM 10 gene.



Result: Search for mammalian promoters (PromoterInspector)

PromoterInspector Release 1.0 **Tue Nov 19 10:41:**

Solution Parameters:

Sequence file: adam_10_genomic.seq (24250 bps)

Inspecting sequence adam_10_genomic (1 - 24250):

Start	End	Length in bp
4841	5836	996

Found 1 promoter regions in 1 sequences

A total of 24250 basepairs was scanned

Please expect PromoterInspector to find good matches but not all Promoters !

- To extract the sequence in between the coordinates 4841 & 5836 I suggest pasting the entire sequence into WORD & using the line numbering function. Since each line has 50 letters you can divide the coordinates by 50 to determine the lines on which your promoter sequence is located. The promoter sequence can then be pasted into MatInspector to look for transcription factor binding sites.

MatInspector professional: Identification of transcription factor binding sites
http://genomatix.de/cgi-bin/matinspector/matinspector.pl

MatInspector professional is a tool that utilizes a library of matrix descriptions for transcription factor binding sites to locate matches in sequences of unlimited length. A large library of predefined matrix descriptions for protein binding sites exists and

has been tested for accuracy and suitability. Similar and/or related matrices have been grouped into matrix families. The matrix library contains 381 weight matrices in five species groups (fungi, insects, plants, vertebrates, and miscellaneous).

Transcription factor binding sites (TF-sites)

Individual TF-sites build the basis of the promoter. These are relatively short stretches of DNA (10 - 20 nucleotides), sufficiently conserved in sequence to allow specific recognition by the corresponding transcription factor. TF-acquisition by DNA binding is the sole function of a TF-site! TF-sites are generally best described by nucleotide weight matrices. MatInspector professional is a good tool for detection of TF-sites in DNA sequences and benefits from a large library of precompiled and quality checked nucleotide weight matrices.

Using MatInspector professional:

Once you have identified the promoter region using the PromoterInspector program you can then search for potential transcription factor binding sites.

- Go to the link for the program above.
- Before you can use this program you will have to register and obtain a user name and password. Do this by clicking "Register now!".
- Once you have obtained a user name and password by e-mail you can use MatInspector professional.
- Click on "start MatInspector professional directly!" and login.
- Enter your sequence in the box provided in one of the supported formats.
- You do not need to change any of the other parameters until the "Library Selection" section.
- Here you can choose whether to search for transcription factor binding sites that have pre-compiled weight matrices (i.e. TFs recognised by the program) **or** you can search your own string of letters by choosing "User-defined IUPAC string (IUPAC)". You can also cut your sequence with a variety of restriction enzymes but we will not be dealing with this feature here.
- Only the IUPAC symbols ABCDGHKMNRSTUVWY can be used (e.g. R is A or G), all other letters are ignored. Specify the maximum number of mismatches allowed in matches to the string. The number of mismatches should not exceed 50% of the string-length.

- Click "Continue".
- On the next page you can change matrix and output parameters.
- You can read about these parameters in detail by clicking the [?] at the top of the page.
- Choose "Vertebrates" as the matrix group, change the core similarity to 1.00 (at least for 1st run), sort matches by quality & leave the other defaults as they are.
- Fill in your e-mail address and click "Submit Query".
- The output is shown below. Remember because a transcription factor binding site exists does not mean that it is functional. This program just gives you a starting point to experimentally characterise your promoter sequence.

Example: promoter region for human ADAM 10 gene identified by PromoterInspector. Coordinates 4750-5000bp (TSS @ 5000bp).

Inspecting sequence adam_10_promoter_1 (1 - 250):

Name of family/matrix	Further Information	RE	Opt.	Position	Strand	Core sim.	Matrix sim.	Sequence
V$HAML/AML1.01	runt-factor AML-1	6.80	1.00	9 - 14	(+)	1.000	1.000	tgTGGT
V$CREB/CREBP1CJUN.01	CRE-binding protein 1/c-Jun heterodimer	0.22	0.88	61 - 68	(+)	1.000	1.000	tgACGTta
V$VBPF/VBP.01	PAR-type chicken vitellogenin promoter-binding protein	3.78	0.85	89 - 98	(+)	1.000	0.877	cTTACatctt
V$GATA/GATA3.02	GATA-binding factor 3	2.01	0.90	90 - 99	(-)	1.000	0.904	gaAGATgtaa
V$NFKB/CREL.01	c-Rel	2.74	0.90	121 - 130	(-)	1.000	0.912	aggaccTTCC
V$GATA/LMO2COM.02	complex of Lmo2 bound to Tal-1, E2A proteins, and GATA-1, half-site 2	2.02	0.95	155 - 163	(-)	1.000	0.967	gaGATAgtg
V$EGRF/EGR3.01	early growth response gene 3 product	0.01	0.76	163 - 174	(+)	1.000	0.837	ccGCGTaggagc
V$SP1F/SP1.01	stimulating protein 1 SP1, ubiquitous zinc finger transcription factor	1.70	0.88	175 - 187	(-)	1.000	0.926	ccggGGCGggagc
V$SP1F/SP1.01	stimulating protein 1 SP1, ubiquitous zinc finger transcription factor	1.70	0.88	184 - 196	(+)	1.000	0.944	ccggGGCGggacc
V$SP1F/SP1.01	stimulating protein 1 SP1, ubiquitous zinc finger transcription factor	1.70	0.88	203 - 215	(-)	1.000	0.901	gggaGGCGgggtt
V$ECAT/NFY.01	nuclear factor Y (Y-box binding factor)	0.49	0.89	215 - 230	(+)	1.000	0.894	caagcCCAAtcccagc
V$PCAT/CLTR_CAAT.01	Mammalian C-type LTR CCAAT box	<0.01	0.74	219 - 243	(+)	1.000	0.751	cCCAAtcccagctctccgcc

Other Resources on the web for nucleic acid sequence analysis

There are many resources available on the web for nucleic acid sequence analysis for a starting point take a look at:

Deambulum - http://www.infobiogen.fr/services/deambulum/english/menu.html

Day 2: Protein Sequence Analysis

TOPICS

1. Physico-chemical properties.
2. Cellular localization.
3. Signal peptides.
4. Transmembrane domains.
5. Post-translational modifications.
6. Motifs & domains.
7. Secondary structure.
8. Other resources.

ExPASy - http://www.expasy.ch/

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) protein and proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures. Besides the tools that we will introduce in this manual there are many other applications available at this website that you should take some time to have a look at.

1) Physico-chemical properties:

ProtParam tool – http://www.expasy.ch/tools/protparam.html

Calculates lots of physico-chemical parameters of a protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)

Example: Human BRCA 1
You can paste the gene sequence from the Course Website.

- At ExPASy → "Proteomics Tools" → "Primary structure analysis".
- Click on the "ProtParam" link.
- Paste your sequence in the box provided
- The sequence must be written using the one letter amino acid code:
- Press the "Compute parameters" button.
- The output for this sequence is shown below.

Number of amino acids: 1863
Molecular weight: 207720.8
Theoretical pI: 5.29
Amino acid composition:

Ala (A) 84	4.5%
Arg (R) 76	4.1%
Etc etc	
Thr (T) 111	6.0%
Trp (W) 10	0.5%
Tyr (Y) 31	1.7%
Val (V) 101	5.4%
Asx (B) 0	0.0%
Glx (Z) 0	0.0%
Xaa (X) 0	0.0%

Total number of negatively charged residues (Asp + Glu): 283
Total number of positively charged residues (Arg + Lys): 213
Atomic composition:

Carbon	C	8908
Hydrogen	H	14246
Nitrogen	N	2554
Oxygen	O	3014
Sulfur	S	74

Formula: $C_{8908}H_{14246}N_{2554}O_{3014}S_{74}$
Total number of atoms: 28796

Extinction coefficients:
Conditions: 6.0 M guanidium hydrochloride
0.02 M phosphate buffer
pH 6.5

Extinction coefficients are in units of M^{-1} cm^{-1} .
The first table lists values computed assuming ALL Cys residues appear as half cystines, whereas the second table assumes that NONE do.

	276 nm	278 nm	279 nm	280 nm	282 nm
Ext. coefficient	102140	102194	100935	99220	95840
Abs 0.1% (=1 g/l)	0.492	0.492	0.486	0.478	0.461

	276 nm	278 nm	279 nm	280 nm	282 nm
Ext. coefficient	98950	99400	98295	96580	93200
Abs 0.1% (=1 g/l)	0.476	0.479	0.473	0.465	0.449

Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:
The instability index (II) is computed to be 54.68
This classifies the protein as unstable.

Aliphatic index: 69.01
Grand average of hydropathicity (GRAVY): -0.785

2) Cellular localization:

PSORT - http://psort.nibb.ac.jp/form2.html

PSORT, a program to predict the subcellular localization sites of proteins from their amino acid sequences. This program makes use of the fact that proteins destined for particular subcellular localizations have distinct amino acid properties particularly in their N-terminal regions. These properties can be used to predict whether a protein is localized in the cytoplasm, nucleus, mitochondria, or is retained in the ER, or destined for lysosome (vacuolar) or the peroxisome. There is a detailed page of output that we can probably ignore. At the end of the output the percentage likelihood of the subcellular localization is given. If you want to learn more about the output and how subcellular localization is determined please see the user manual at:
http://psort.nibb.ac.jp/helpwww2.html

Example: Human ETS-1 protein.

- At ExPASy → "Proteomics Tools" → "Post-translational modification prediction".
- Click on the "PSORT" link.
- For animal/yeast sequences click the link to "PSORT II Prediction".
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Press the submit button.
- The output for this sequence is shown below.

There are a number parameters measured by this program which you can read about as links from the output file. By scrolling to the bottom of the output you can see the probability that this sequence is nuclear, cytoplasmic, peroxisomal, vacuolar or cytoskeletal. PSORT predicts that ETS-1 is nuclear with a high probability. The fact that ETS-1 is localized in the nucleus has been previously experimentally determined.

Results of Subprograms

```
PSG:  a new signal peptide prediction method
      N-region:  length 8;  pos.chg 2;  neg.chg 1
      H-region:  length 6;  peak value   1.89
```

```
      PSG score:  -2.51

GvH:  von Heijne's method for signal seq. recognition
      GvH score (threshold: -2.1): -10.14
      possible cleavage site: between 54 and 55

>>> Seems to have no N-terminal signal peptide

ALOM: Klein et al's method for TM region allocation
      Init position for calculation: 1
      Tentative number of TMS(s) for the threshold  0.5:   0
      number of TMS(s) .. fixed
      PERIPHERAL  Likelihood =  3.61 (at 98)
      ALOM score:   3.61  (number of TMSs: 0)

MITDISC: discrimination of mitochondrial targeting seq
      R content:       0       Hyd Moment(75):  6.78
      Hyd Moment(95):  6.47    G content:       0
      D/E content:     2       S/T content:     3
      Score: -6.01

Gavel: prediction of cleavage sites for mitochondrial preseq
      cleavage site motif not found

NUCDISC: discrimination of nuclear localization signals
      pat4: none
      pat7: none
      bipartite: none
      content of basic residues:  11.3%
      NLS Score: -0.47

KDEL: ER retention motif in the C-terminus: none

ER Membrane Retention Signals: none

SKL: peroxisomal targeting signal in the C-terminus: none

SKL2: 2nd peroxisomal targeting signal:  none

VAC: possible vacuolar targeting motif: none

RNA-binding motif: none

Actinin-type actin-binding motif:
      type 1: none
      type 2: none

NMYR: N-myristoylation pattern : none

Prenylation motif: none

memYQRL: transport motif from cell surface to Golgi: none

Tyrosines in the tail: none

Dileucine motif in the tail: none

checking 63 PROSITE DNA binding motifs:
      Ets-domain signature 1 (PS00345):  *** found ***
         LWQFLLELL at 337

      Ets-domain signature 2 (PS00346):  *** found ***
         KPKMNYEKLSRGLRYY at 381

checking 71 PROSITE ribosomal protein motifs:  none

checking 33 PROSITE prokaryotic DNA binding motifs:  none

NNCN: Reinhardt's method for Cytplasmic/Nuclear discrimination
      Prediction: nuclear
      Reliability: 55.5

COIL: Lupas's algorithm to detect coiled-coil regions
```

```
      total: 0 residues
```

```
Results of the k-NN Prediction
k = 9/23
          73.9 %: nuclear
          13.0 %: cytoplasmic
           4.3 %: peroxisomal
           4.3 %: vacuolar
           4.3 %: cytoskeletal

>> prediction for QUERY is nuc (k=23)
```

3) Signal peptides:

Proteins destined for secretion, operation with the endoplasmic reticulum, lysosomes and many transmembrane proteins are synthesized with leading (N-terminal) 13 – 36 residue signal peptides.

SignalP - http://www.cbs.dtu.dk/services/SignalP/

The SignalP WWW server can be used to predict the presence and location of signal peptide cleavage sites in your proteins. It can be useful to know whether your protein has a signal peptide as it indicates that it may be secreted from the cell. Furthermore, proteins in their active form will have their signal peptides removed, if you can determine the length of the signal peptide then you can calculate the size of the protein minus the signal peptide.

Example: Human Beta-defensin; sp|Q09753|BD01_HUMAN

- At ExPASy → "Proteomics Tools" → "Post-translational modification prediction".
- Click on the "SignalP" link.
- Scroll to bottom of page.
- Paste your sequence in the box provided
- The sequence must be written using the one letter amino acid code:
- It is recommend that the N-terminal part only (not more than 50-70 amino acids) of the sequences is submitted. A longer sequence will increase the risk of false positives and make the graphical output difficult to read.
- Choose one or more group of organisms for the prediction by clicking the check-box next to the group(s):

gram-: Use networks trained on sequences from gram-negative prokaryotes
gram+: Use networks trained on sequences from gram-positive prokaryotes
euk: Use networks trained on sequences from eukaryotes

- If no groups are indicated, predictions from all three groups will be returned.
- A graphical output (in Postscript format) of the prediction will be available, if the "Include graphics"-button is checked.
- Press the "Submit sequence" button.

A WWW page will return the results when the prediction is ready. Response time depends on system load. The output for this sequence is shown below

C score = raw cleavage site score

The output score from networks trained to recognize cleavage sites vs. other sequence positions. Trained to be: High at position +1 after the cleavage site and low at all other positions.

S score = signal peptide score

The output score from networks trained to recognize signal peptide vs. non-signal-peptide positions. Trained to be: High at position before the cleavage site and low at all other positions.

Y score = combined cleavage site score

The prediction of cleavage site location is optimized by observing where the C-score is high and the S-score changes from a high to a low value.

For each sequence, SignalP will report the maximal C, S, and Y scores, and the mean S-score between the N-terminal and the predicted cleavage site. These values are used to distinguish between signal peptides and non-signal peptides. If your sequence is predicted to have a signal peptide, the cleavage site is predicted to be immediately before the position with the maximal Y-score.

The Human beta-defensin protein has a predicted signal peptide from position 1 to 21 and a potential cleavage site exists between positions 21 and 22. These predictions correspond exactly to the SWISS-PROT annotation for this protein (accession Q09753).

SignalP V1.1 World Wide Web Server

- Explanation of the output

********** SignalP predictions **********
Using networks trained on euk data

>Sequence length = 68

```
# pos  aa    C       S       Y
    1   M   0.012   0.967   0.009
    2   R   0.012   0.965   0.010
    3   T   0.014   0.965   0.014
    4   S   0.014   0.952   0.021
    5   Y   0.012   0.956   0.027
    6   L   0.012   0.956   0.035
    7   L   0.012   0.949   0.043
    8   L   0.012   0.965   0.048

etc etc

   65   K   0.069   0.055   0.038
   66   C   0.013   0.049   0.017
   67   C   0.023   0.056   0.022
   68   K   0.018   0.053   0.019

< Is the sequence a signal peptide?
# Measure  Position  Value  Cutoff  Conclusion
  max. C    22       0.848   0.37   YES
  max. Y    22       0.832   0.34   YES
  max. S    12       0.983   0.88   YES
  mean S     1-21    0.915   0.48   YES
# Most likely cleavage site between pos. 21 and 22: ASG-GN
```

Please cite:
Henrik Nielsen, Jacob Engelbrecht, Søren Brunak and Gunnar von Heijne: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering, **10**, 1-6 (1997).

4) Transmembrane domains:

Tmpred - http://www.ch.embnet.org/software/TMPRED_form.html

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. The presence of transmembrane domains is an indication that the protein is located on the cell surface.

Example: Human chemokine receptor 4 protein sequence NP_003458.1

- At ExPASy → "Proteomics Tools" → "transmembrane region detection".
- Click on the link to Tmpred.
- Paste your sequence in the box provided in one of the supported formats e.g.
- plain text, SwissProt_ID or AC, etc.
- You may change the minimal and maximal length of the hydrophic part of the
- transmembrane helix but unless you have reason to do so you should accept the defaults i.e. 17 and 33. ~22 residues is the same length as the width of a lipd bilayer.
- Click the "Run Tmpred" button to start the search.
- The output is given in 3 parts 1, 2 and 3 (see below).
- Part 1: lists all the significant predictions of possible transmembrane helices
- in this case there are 7 helices predicted but at this stage we do not know the orientation of the helices so there are 2 tables, the first with the helices orientated from the inside to the outside and vice versa for the second.
- Part 2: shows which inside->outside helices correspond to the outside -> inside helices and indicates which orientation is most likely.
- Part 3: proposes the strongly preferred model for the transmembrane domain
- structure of the protein and also an alternative model.
- A graphic of the prediction is also available (not shown here)
- These predictions correspond well but not exactly to the SWISS-PROT
- annotation for this protein (accession P30991).

Tmpred output
[ISREC-Server] Date: Mon Dec 10 13:11:02 MET 2001

Sequence: MEG...HSS, length: 352
Prediction parameters: TM-helix length between 17 and 33

---

1. Possible transmembrane helices
The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

```
Inside to outside helices :   7 found

      from        to    score center
  39 (  46)  62 (  62)   1962     54
  78 (  85) 105 ( 103)   1623     95
 114 ( 114) 133 ( 130)   1352    122
 155 ( 157) 175 ( 173)   1716    165
 204 ( 206) 223 ( 223)   2052    214
 240 ( 240) 261 ( 259)   2840    251
 286 ( 286) 305 ( 305)   1241    295

Outside to inside helices :   7 found
      from        to    score center
  47 (  47)  63 (  63)   2568     55
  78 (  78)  96 (  96)   1331     86
 111 ( 114) 132 ( 132)   1740    122
 155 ( 157) 173 ( 173)   1197    165
 204 ( 204) 223 ( 223)   2404    214
 240 ( 242) 259 ( 259)   2037    251
 283 ( 286) 305 ( 305)   1703    294
```

---

2. Table of correspondences
Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.
Helices shown in brackets are considered insignificant.
A “+”-symbol indicates a preference of this orientation.
A “++”-symbol indicates a strong preference of this orientation.

```
          Inside->outside | outside->inside
    39-  62 (24) 1962      |    47-  63 (17) 2568 ++
    78- 105 (28) 1623 ++   |    78-  96 (19) 1331
   114- 133 (20) 1352      |   111- 132 (22) 1740 ++
   155- 175 (21) 1716 ++   |   155- 173 (19) 1197
   204- 223 (20) 2052      |   204- 223 (20) 2404 ++
   240- 261 (22) 2840 ++   |   240- 259 (20) 2037
   286- 305 (20) 1241      |   283- 305 (23) 1703 ++
```

---

3. Suggested models for transmembrane topology
These suggestions are purely speculative and should be used with **extreme caution** since they are based on the assumption that all transmembrane helices have been found.
In most cases, the Correspondence Table shown above or the prediction plot that is also created should be used for the topology assignment of unknown proteins.

2 possible models considered, only significant TM-segments used

---→ STRONGLY preferred model: N-terminus outside

```
 7 strong transmembrane helices, total score : 14594
 # from   to length score orientation
 1   47   63 (17)    2568 o-I
 2   78  105 (28)    1623 I-o
 3  111  132 (22)    1740 o-I
 4  155  175 (21)    1716 I-o
 5  204  223 (20)    2404 o-I
 6  240  261 (22)    2840 I-o
 7  283  305 (23)    1703 o-I

----→ alternative model
 7 strong transmembrane helices, total score : 11172
 # from   to length score orientation
 1   39   62 (24)    1962 I-o
 2   78   96 (19)    1331 o-I
 3  114  133 (20)    1352 I-o
 4  155  173 (19)    1197 o-I
 5  204  223 (20)    2052 I-o
 6  240  259 (20)    2037 o-I
 7  286  305 (20)    1241 I-o
```

5) Post-translational modifications:

After translation has occurred proteins may undergo a number of posttranslational modifications. These can include the cleavage of the pro- region to release the active protein, the removal of the signal peptide and numerous covalent modifications such as, acetylations, glycosylations, hydroxylations, methylations and phosphorylations. Posttranslational modifications such as this may alter the molecular weight of your protein and thus its position on a gel. There are many programs available for predicting the presence of posttranslational modifications, we will take a look at one for the prediction of type O-glycosylation sites in mammalian proteins. Remember these programs work by looking for consensus sites and just because a site is found does not mean that a modification definitely occurs.

NetOGlyc - http://www.cbs.dtu.dk/services/NetOGlyc/

Prediction of type O-glycosylation sites in mammalian proteins. This program works by comparing the input sequence to a database of 299 known and verified mucin type O-glycosylation sites extracted from O-GLYCBASE.

Example: Human CD1D sp|P15813|CD1D_HUMAN

- At ExPASy → "Proteomics Tools" → "Post-translational modification".
- Click on the link to "NetOGlyc".
- Paste your sequence in the box provided in FASTA format.
- Check "generate graphics" and click the submit button.
- The output for this program is shown below (graphics not shown).
- This program predicts potential O-glycosylation sites at Threonine 64 and Serine 214.

NetOGlyc 2.0 Prediction Results

```
Name:  Sequence      Length:  335
MGCLLFLLLWALLQAWGSAEVPQRLFPLRCLQISSFANSSWTRTDGLAWLGELQTHSWSNDSDTVRSLKPWSQGTFSDQQ      80
WETLQHIFRVYRSSFTRDVKEFAKMLRLSYPLELQVSAGCEVHPGNASNNFFHVAFQGKDILSFQGTSWEPTQEAPLWVN     160
LAIQVLNQDKWTRETVQWLLNGTCPQFVSGLLESGKSELKKQVKPKAWLSRGPSPGPGRLLLVCHVSGFYPKPVWVKWMR     240
GEQEQQGTQPGDILPNADETWYLRATLDVVAGEAAGLSCRVKHSSLEGQDIVLYWGGSYTSMGLIALAVLACLLFLLIVG     320
FTSRFKRQTSYQGVL
............................................................T.................      80
................................................................................     160
................................................S...............................     240
................................................................................     320
...............
```

Name	Residue	No.	Potential	Threshold	Assignment
Sequence	Thr	42	0.0611	0.6493	.
Sequence	Thr	44	0.0087	0.6573	.
Sequence	Thr	55	0.0117	0.6491	.
Etc etc					
Sequence	Thr	248	0.0131	0.5840	.
Sequence	Thr	260	0.0089	0.6578	.
Sequence	Thr	266	0.0224	0.6957	.
Sequence	Thr	300	0.0147	0.7357	.
Sequence	Thr	322	0.0480	0.7096	.
Sequence	Thr	329	0.0639	0.6021	.

Name	Residue	No.	Potential	Threshold	Assignment
Sequence	Ser	18	0.0161	0.6211	.
Sequence	Ser	34	0.0044	0.6673	.
Sequence	Ser	35	0.0048	0.6717	.
Sequence	Ser	39	0.0337	0.6118	.
Sequence	Ser	40	0.0013	0.6521	.
Sequence	Ser	57	0.0065	0.6484	.
Etc etc					
Sequence	Ser	284	0.0005	0.6401	.
Sequence	Ser	285	0.0082	0.6389	.
Sequence	Ser	298	0.0003	0.6778	.
Sequence	Ser	301	0.0007	0.6924	.
Sequence	Ser	323	0.0003	0.6441	.
Sequence	Ser	330	0.0052	0.6277	.

6) Motifs and Domains

If you want to determine the function of a protein the first tool of choice is homology searching (see day 4). Unless this finds you a match with a well charaterised protein comprehending the entire length of yours you should look for motifs and domains in your protein.To determine if your protein sequence contains known motifs or conserved domain structures you should search the protein against one of the motif or profile databases. There are many of these available but we will discuss ProfileScan, which allows you to search both the Prosite and Pfam databases simultaneously. See the documentation for more details.

ProfileScan - http://hits.isb-sib.ch/cgi-bin/PFSCAN

Example: Human CFTR sp|P13569|CFTR_HUMAN

- Go to the URL above
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Tick the motif databases you wish to search, other parameters should be OK.
- Press the "scan" button.

The output for this program is too large to show here, but it gives lots of detail about motifs in the CFTR protein identifying potential: ABC transporters family signature; ATP/GTP-binding site motif A (P-loop); Protein kinase C phosphorylation sites; N-glycosylation sites; Casein kinase II phosphorylation site; N-myristoylation sites; cAMP- and cGMP-dependent protein kinase phosphorylation site; Bipartite nuclear localization signal; NACHT-NTPase domain profile; Guanylate kinase domain profile etc.

Remember that these programs only tell you are that there is a motif present and thus there is the potential for these modifications and functions to occur. It is up to you to determine experimentally which are real but at least you now know what to look for.

7) Secondary Structure Prediction

If protein structure, even secondary structure, can be accurately predicted from the now abundantly available gene and protein sequences, such sequences become immensely more valuable for the understanding of drug- design, the genetic basis of disease, the role of protein structure in its enzymatic, structural, and signal transduction functions, and basic physiology from molecular to cellular, to fully systemic levels. In short, the solution of the protein structure prediction problem (and the related protein folding problem) will bring on the second phase of the molecular biology revolution (Munson *et al*., 1994).

JPRED - http://www.compbio.dundee.ac.uk/~www-jpred/submit.html

Jpred is an Internet web server that takes either a protein sequence or a multiple alignment of protein sequences, and predicts secondary structure. It works by combining a number of modern, high quality prediction methods to form a consensus. Please be aware that secondary structure prediction is an extremely complex problem that is under intensive research and we are still at a relatively primitive stage.

We cannot discuss the details of protein secondary structure here but if you are interested in this area we recommend that you take a look at any major biochemistry textbook.

Essentially protein secondary structure consists of 3 major conformations; the α Helix, the β pleated sheet and the coil conformation.

Example: Human alpha 1 hemoglobin. NP_000549.1

- At the ExPASy → "Proteomics Tools" → "Secondary structure prediction".
- Click on the link to JPRED.
- Click "Prediction".
- Paste your sequence in the box provided.
- The defaults are OK.
- Click "Run secondary structure predictions!"
- Point 5 on the submission page allows you to deselect the BLAST search against PDB (Protein Data Bank). If your sequence already has had its structure predicted or experimentally determined it will be in here and you can follow the link to PDB for information on the structure of your protein.

- If your protein is in PDB you can view your protein secondary structure using RasMol (To download RasMol see the course website for a link)
- Once you have RasMol running you can open your structure in it a view it using a number of different options.

Otherwise continue with prediction

- The program may take a long time so you can save a bookmark and return to your results later or choose to have your results e-mailed to you.
- There are a number of options to view the output, view your output in HTML format (option 4).
- The complete output is too large to show here (see webpage).
- Scroll down through the output until you get to "Jpred" output. The line of output beside this is the consensus secondary structure for your sequence. H= Helices E= strands C= coils.

A Few Other Useful Tools at ExPASy

FindMod

Predicts potential protein post-translational modifications (PTM) and find potential single amino acid substitutions in peptides. The experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SWISS-PROT/TrEMBL entry or from a user-entered sequence, and mass differences are used to better characterise the protein of interest.

NetPhos:

The NetPhos WWW server produces neural network predictions for serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins.

Sulfinator:

Predicts tyrosine sulfation sites in protein sequences. Tyrosine sulfation is an important post-translational modification of proteins that go through the secretory pathway.

REP:

Searches a protein sequence for a collection of repeats such as leucine rich repeats and many others.

Other Resources for Protein Sequence Analysis

1) Protein Prospector at UCSF - http://prospector.ucsf.edu/

MS-Digest: A protein digestion tool from the UCSF Mass Spectrometry Facility that performs an *in silico* enzymatic digestion of a protein sequence, and calculates the mass of each peptide.

MS-Product: A tool from the UCSF Mass Spectrometry Facility that calculates the possible fragment ions resulting from fragmentation of a peptide in a mass spectrometer. Fragmentation possibilities for post-source decay (PSD), high-energy collision-induced dissociation (CID), and low-energy CID processes may be calculated.

2) Pasteur Institute - http://bioweb.pasteur.fr/seqanal/protein/intro-uk.html

Antigenic: finds antigenic sites in proteins.

Helixturnhelix: reports nucleic acid binding motifs in your protein of interest.

DAY 3: Accessing Completed Genomes

TOPICS:

1. **UCSC Genome Bioinformatics -** Human & Mouse Genomes.

2. **Ensembl** – Human, Mouse, Zebrafish, Fugu & Mosquito Genomes.

3. **NCBI Genomic Biology** – Human, Mouse, Rat, Zebrafish, Drosophila, Malaria, Plants, Microbes, Retroviruses.

Accessing Genomic Sequences:

There is no one resource available on the web that allows you to access all the available genomes. In this course we will take a look at 3 excellent sites for accessing most of the genomic information that is available out there – UCSC Genome Bioinformatics; Ensembl & NCBI Genomic Biology. These sites often contain similar information and it may be possible to get most of the information you require from just one of these sites, however, to get the maximum amount of information it is often worth having a look at all 3 of these sites. In this course we will primarily concentrate on accessing the human genome, however, any of the examples that we describe can easily be applied to any of the available species. Remember that most of the genomes are still in a draft state and are subject to change as more sequence becomes available.

http://genome.cse.ucsc.edu/

At this site the latest assembly of the human and mouse genomes can be accessed. You can choose which one you want to access by using the pull down menu under "Genome".

Once you have decided what genome you want to access there are two major ways to do so – 1) BLAT Search 2) Genome Browser.

BLAT Search:

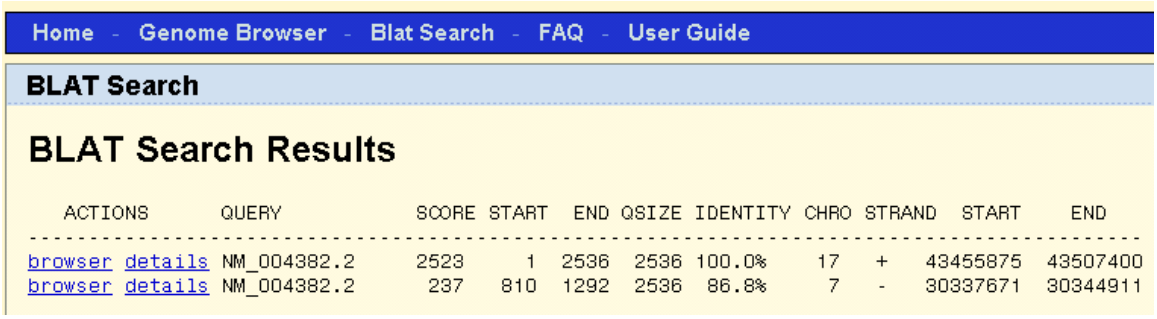
Not to be confused with BLAST, a BLAT search is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more on the genome. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. You can use this tool to locate any DNA/RNA sequence on the genome.

To do a BLAT search:

- Click on "Blat" in the left hand side menu.
- Paste your sequence in the box provided or upload a file containing your sequence using the "Browse…" button.
- Multiple sequences can be searched at once if separated by a line starting with > and the sequence name. (Fasta format)

- Using the pull-down menus choose the genome and assembly you wish to search (default is most recent assembly).
- You can leave the defaults in the other menus as they are (unless you want to search a protein => change "Query type:" to protein) and "Submit".
- This will take you to BLAT Search Results – There may be more than one hit against the genome, but the best hit will be identified by its percentage identity.

Example - Homo sapiens corticotropin releasing hormone receptor 1 (CRHR1) mRNA. [NM_004382.2]


Home - Genome Browser - Blat Search - FAQ - User Guide
BLAT Search
BLAT Search Results
ACTIONS QUERY SCORE START END QSIZE IDENTITY CHRO STRAND START END
browser details NM_004382.2 2523 1 2536 2536 100.0% 17 + 43455875 43507400
browser details NM_004382.2 237 810 1292 2536 86.8% 7 - 30337671 30344911

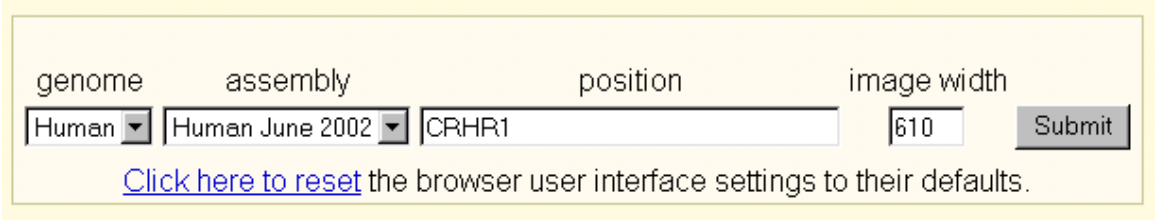
- You can click on either "browser" (see next section) or "details"
- Details – alignment of the mRNA to the genomic sequence. Gives you the intron-exon structure of your gene.

Browser:

The genome can also be accessed via the browser, which is a graphical display of the genome where various features can be displayed at once.

To access the genome via the browser:

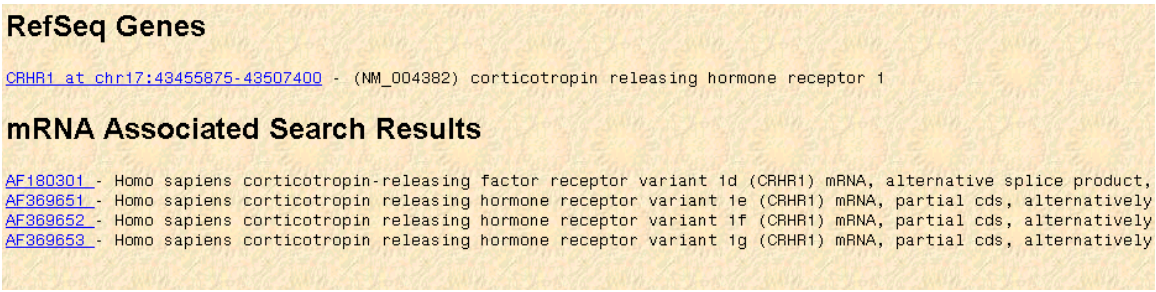
- Click on "Browser" in the left hand side menu of the start page or via BLAT as described above.
- This will bring you to the Genome Browser Gateway.
- Here again you can choose which genome (Human/Mouse) and assembly you wish to access.
- In the "position" box you can enter a number of terms to access a particular region of the genome.



- You can also enter the accession number of a sequenced human genomic clone, an mRNA or EST accession, the name of a fingerprint map contig, an STS marker, a cytological band, a range of a chromosome, or words from the Genbank description of an mRNA such as the gene name.

Example - Homo sapiens corticotropin releasing hormone receptor 1 (CRHR1) mRNA. [NM_004382.2]

- One way to search for this gene is to type CRHR1 in the position box and click "Submit".

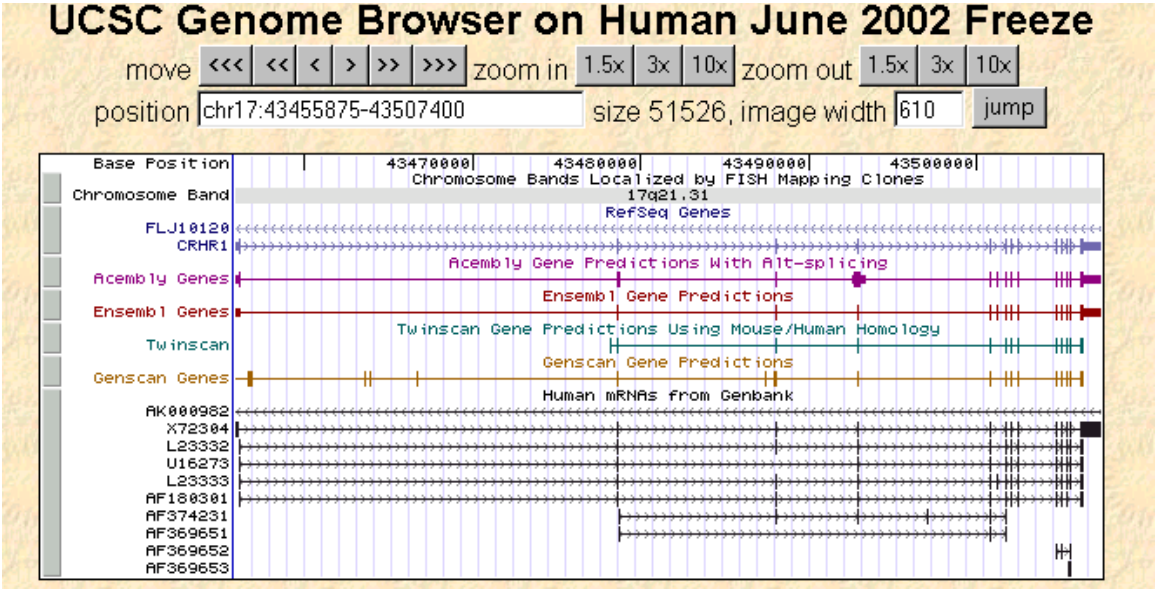

RefSeq Genes
CRHR1 at chr17:43455875-43507400 - (NM_004382) corticotropin releasing hormone receptor 1
mRNA Associated Search Results
AF180301 - Homo sapiens corticotropin-releasing factor receptor variant 1d (CRHR1) mRNA, alternative splice product,
AF369651 - Homo sapiens corticotropin releasing hormone receptor variant 1e (CRHR1) mRNA, partial cds, alternatively
AF369652 - Homo sapiens corticotropin releasing hormone receptor variant 1f (CRHR1) mRNA, partial cds, alternatively
AF369653 - Homo sapiens corticotropin releasing hormone receptor variant 1g (CRHR1) mRNA, partial cds, alternatively

- RefSeq Genes: CRHR1 is a known RefSeq gene (RefSeq is an NCBI database of annotated genes with 1 reference sequence given for any 1 gene) and is located on chromosome 17 at the position shown above.

- mRNA Associated Search Results: Displays the known mRNAs for CRHR1.

- Click on one of the links to take you to a graphical display of the CRHR1 on the genome (see below).

- You can use the zoom buttons to zoom in or out of the current location on the genome enabling you to view a wider or more specific genomic context around your gene. You can also use the move buttons to move along the genome.

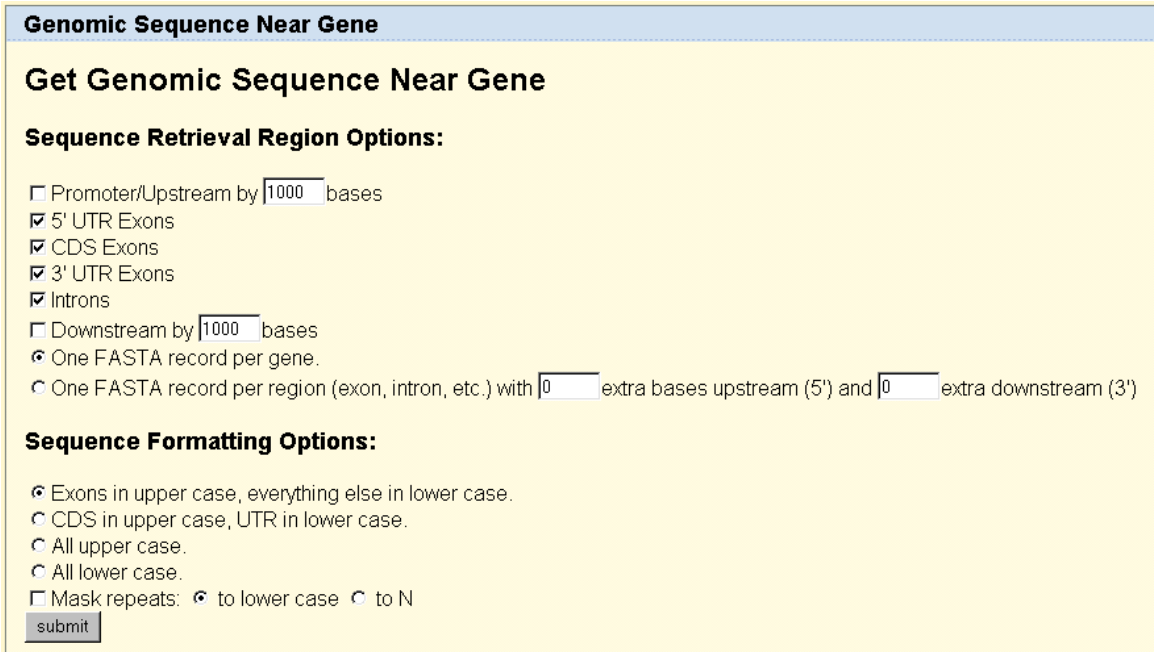


- There are a number of features displayed
  - Base position – the coordinates of the gene on the chromosome.
  - Chromosome band i.e. 17q21.31
  - RefSeq Genes: Known genes in this area – click on one of the links to the left to get more details.
  - Acembly, Ensembl, Twinscan, Genscan Genes are all gene predictions from various computer programs.
  - Human mRNAs from Genbank.

- Click on any of the links for more details.
- Below the graphical display there are a number of other items that you can also choose to display on the browser.
- You can choose to hide these options or display them in "dense" format or "full" format. The full option displays each item on its own line on the browser.
- You can find out about any of the options by clicking on the blue hyperlinks.
- Once you have chosen which options you wish to display click the "refresh" button".

Obtaining Genomic Sequence From UCSC Genome Browser:

- Click anywhere on the RefSeq track. This takes you to a page with information about your gene including links to RefSeq, OMIM, LocusLink, PubMed, GeneLynx, GeneCards, Mouse Ortholog etc (see later sections for details)
- You can follow any of these links to more information on your gene.
- To obtain sequence information click on:
    - Predicted protein
    - mRNA sequence
    - Genomic Sequence
    - Comparative Sequence (alignment to mouse genome)
- Click on the link to "genomic sequence".

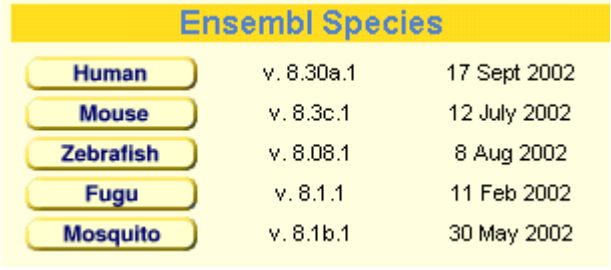

Genomic Sequence Near Gene
Get Genomic Sequence Near Gene
Sequence Retrieval Region Options:
Promoter/Upstream by 1000 bases
5' UTR Exons
CDS Exons
3' UTR Exons
Introns
Downstream by 1000 bases
One FASTA record per gene.
One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
Sequence Formatting Options:
Exons in upper case, everything else in lower case.
CDS in upper case, UTR in lower case.
All upper case.
All lower case.
Mask repeats: to lower case to N
submit

- Tick any of options that you require and then click submit to obtain your sequence.

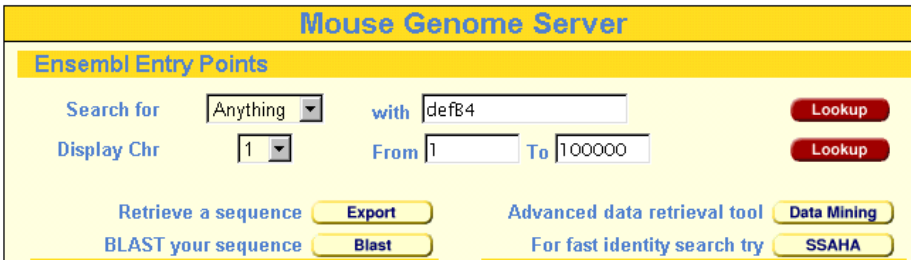
For more detailed information on the UCSC Genome Browser click on the link to the User Guide at the start page.

http://www.ensembl.org

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. To date the human, mouse, zebrafish, Fugu, and mosquito genomes are available.



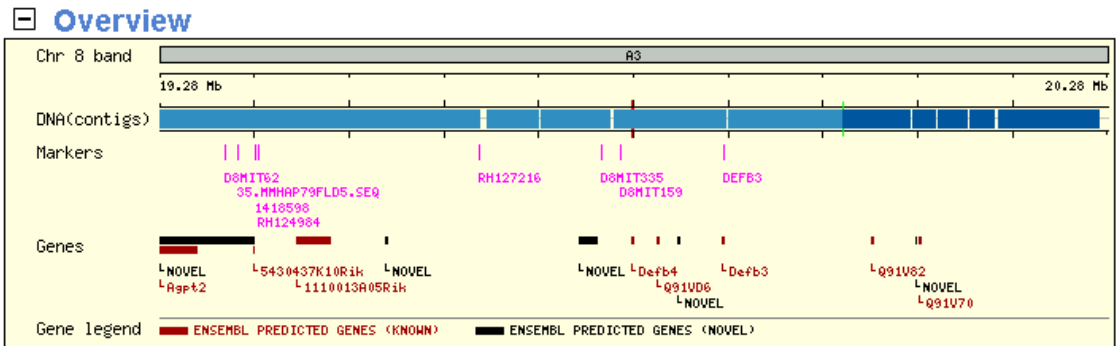
- Click on one of the species to access the genomic information e.g. Mouse.



- To find your gene of interest you can enter in the empty box, the gene symbol, gene accession number, mRNA accession number, SwissProt accession number, EST accession number etc.
- You can also access the genome by chromosome number.
- Below this there are 4 buttons
    - Export – If you have an Ensembl I.D. for your gene you can download its sequence from here.
    - Blast – BLAST your sequence against the genome.
    - Data Mining – Allows the download a large datasets e.g. all the genes on a chromosome, the entire genome etc.
    - SSAHA – Match a DNA sequence to the genomic sequence (similar to BLAT).

Example: Mouse beta-defensin 4 (defB4).

- Use the pull-down "Anything" menu to select "Gene" and type the gene RefSeq symbol in the empty box (defB4).
- Click "Lookup".
- This will take you to a query results page. In this case there is only one hit but sometimes you will have to look through a number of entries to find what you are looking for.
- Click on the **EnsEMBL Gene:** ENSMUSG00000031470 link for information on your gene such as it's sequence, structure, domains that it contains etc.
- Click on the link to "Genomic Location" to display the gene in the genome similar to the UCSC browser.
- There are 3 major views displayed
    - Chromosome – highlights position on the chromosome
    - Overview – shows genes surrounding gene of interest on chromosome band



- Detailed View – more detailed view of your gene.

- On the detailed view you can use the pull down menus to "Features", "DAS Sources" & decorations to choose what details you wish to display (similar to UCSC)
- The "Export" menu can be used to download the genomic sequence or features of this area in the genome (E.g. list of genes).
- As with the UCSC browser you can zoom in or out of this region of genome or move along the genome using the "window" buttons.

http://www.ncbi.nlm.nih.gov/Genomes/index.html

Many of you will be familiar with the National Center for Biotechnology Information (NCBI) website which has many very useful resources including; Entrez, PubMed, Genbank, BLAST, OMIM. Today we will see how to use the NCBI site to interrogate the genomic sequences that are available there.
The NCBI site provides a good starting point for accessing the human, mouse, Rat, Zebrafish, Drosophila, Malaria, Plant, microbial and viral genomes. Many of these genomes will have their own dedicated sites located at other websites, but the NCBI site will provide links to them.

Accessing The Human Genome:

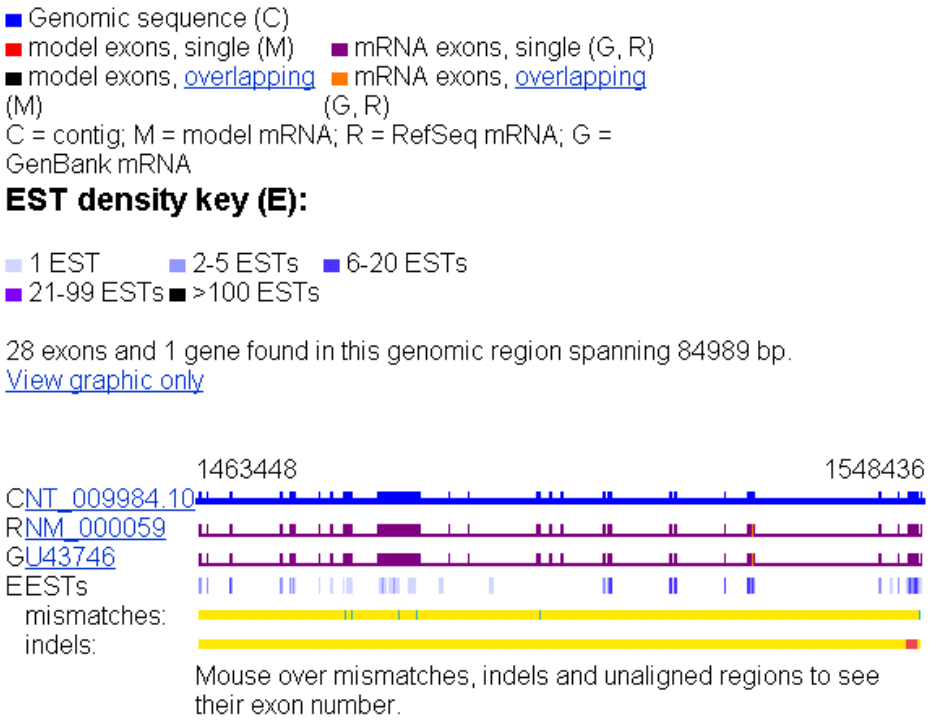
To access the human genome - go to the URL above and click on the link to "Human". This page provides a number of links such as a link to BLAST where you can search your sequence against the human genome. You can also browse the genome by chromosome by clicking on one of the chromosomes. The best way to access the genome if you have a particular gene of interest is to search for your gene in LocusLink. LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.

- At the top of the page search LocusLink by entering your gene name (full name, abbreviation or accession number) in the box and "Go". (Example: BRCA2)
- This brings up a results page that matches the query for some reason.
- Click on LocusID (e.g. 675) corresponding to the gene to take you to the LocusLink page for that gene.

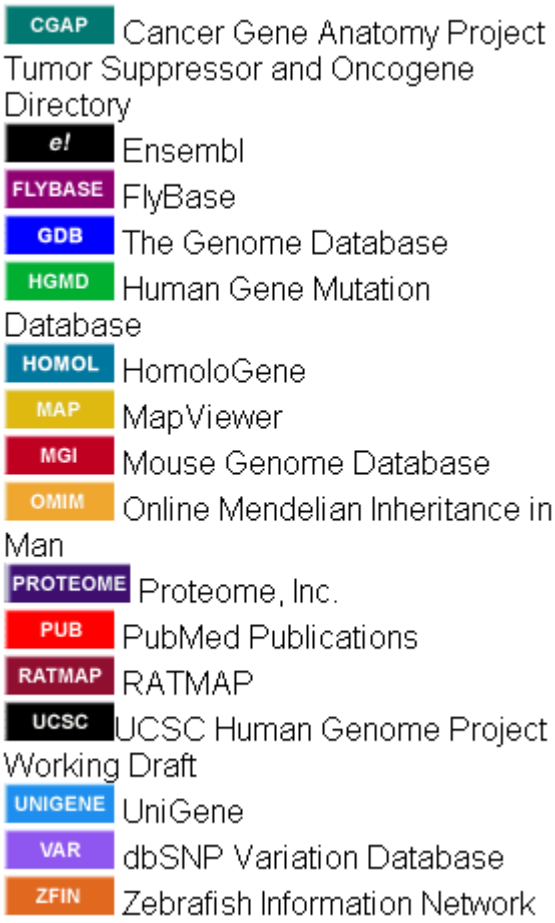
Starting at the top of the page:

- To view to intron-exon structure and sequence of the gene follow the link "Click to Display mRNA-Genomic Alignments". Then click on "Go to full display with alignments".

- Shown below are the known mRNAs and ESTs mapped to the genomic sequence. BRCA2 has 28 exons! By scrolling down the page you can obtain the actual genomic sequence.



- Back at the LocusLink page there are a number of coloured buttons that provide outside links to other sites with information on the gene of interest. Some or all of these may be present depending on the particular gene you search.



Moving Down the Page:

- **Official Gene Symbol** - The official symbol and full name of a gene assigned by the genome-specific nomenclature committee.
- **Interim Gene Symbol** - If the official symbol and name have NOT been established, an interim symbol and name are provided.
- **LocusID** - The unique NCBI LocusID assigned to each locus. LocusIDs are stable, genome-independent identifiers.
- **RefSeq Summary** - written by staff of the NCBI RefSeq group describing the function, localization, and sequence properties of the gene and its products.
- **Type** - The category of the locus. Options include genes that encode proteins, genes that encode untranslated RNA, mapped phenotypes, anonymous DNA segments, models.

- **Product** - The preferred gene product name. These names may be revised to reflect current usage.
- **Alternate Symbols** - Alias symbols. These are compiled from previous nomenclature, the published literature, or sequence records.
- **Alias** - Alternative product names.
- **Function -** This section describes the function of the protein or RNA encoded by this locus. It may include Enzyme Commission numbers, disease names established by OMIM, and/or terms from ontologies such as Gene Ontology (GO).
- **Relationships** - This section is under development. At present it may report human/mouse homologies and transcripts annotated on the genome (models) that are related to known genes.
- **Map Information:**
    - **Chromosome** - The chromosome(s) to which this locus has been mapped.
    - **Cytogenetic** - The cytogenetic position.
    - **Genetic** - The genetic position.
    - **Markers STS** - markers determined to be associated with this locus. These associations are established either computationally by ePCR with mRNAs associated with this locus_id, or computationally from the NCBI's Genome Annotation based on STS lying within the region defined by a LocusID, or in a published citation.

- NCBI Reference Sequences (RefSeq)All RefSeq records created for a given locus are listed. Multiple records are distinguished by the brief description of the transcript variant. This section provides links to:
    - RefSeq nucleotide record (genomic and mRNA accessions have 'NG_' and 'NM_' prefixes, respectively)

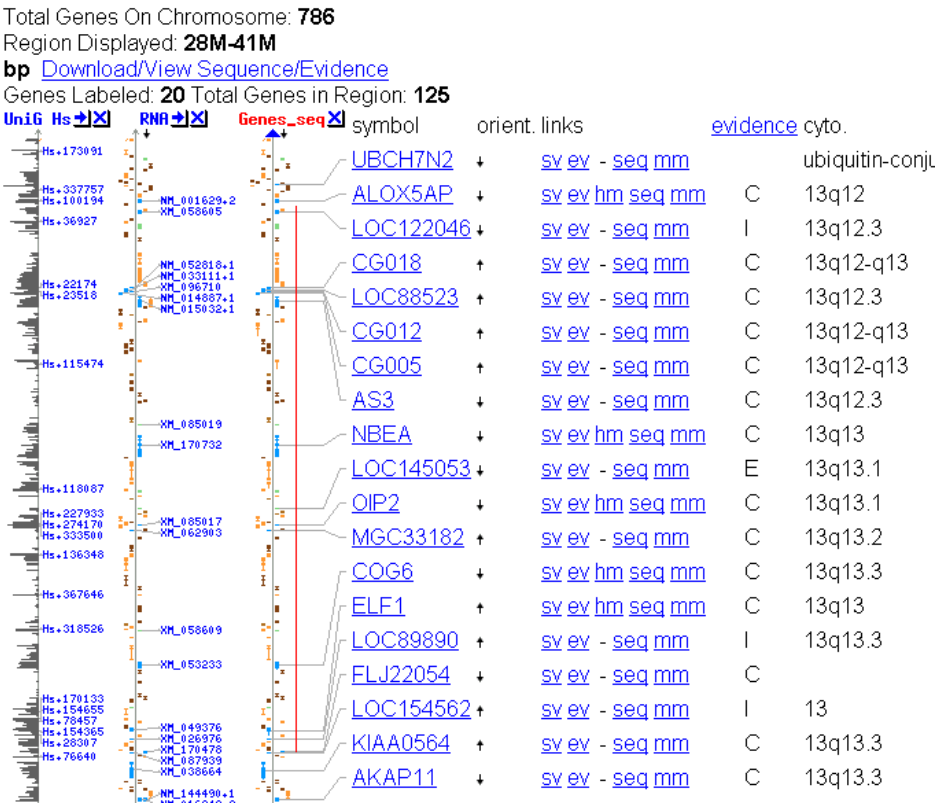
- RefSeq protein record (the 'NP_' prefix)
- CD browser report for each domain found in the RefSeq
- GenBank source sequence(s) used to construct the RefSeq record.
- BL provides a quick link to the BLink result for the protein.

- **GenBank Sequences** - A table of a subset of representative nucleotide and protein accessions for the locus. EST accession numbers are provided if no other sequence data are available to represent the locus (m = mRNA; g = genomic DNA; e = EST). Protein accessions are linked to Entrez sequence displays – Click "BL".
- **Additional Links** - This section names and provides links to additional sites that may contain information related to this locus such as OMIM, UniGene, etc.

MAPVIEW:

This is the NCBI graphical display tool, which you can use to display the genomic context of your sequence. This tool is not as user friendly or as advanced as the UCSC or Ensembl browsers and we recommend that you use these to view the genome graphically where possible. Not all species are available at these sites so you may need to use Mapview.

- To access Mapview click the "mv" links that occur throughout the LocusLink record.
- Click on "Maps & Options" to choose which features you wish to display (here I have chosen to display Unigenes, RNAs and known genes).
- Click on any of the genes, RNAs or Unigenes to get more information.
- You can download genomic sequence for the region selected using the "Download/View Sequence/Evidence" link.


Total Genes On Chromosome: 786
Region Displayed: 28M-41M
bp Download/View Sequence/Evidence
Genes Labeled: 20 Total Genes in Region: 125
UniG Hs
RNA
Genes_seq
symbol orient. links evidence cyto.
UBCH7N2 sv ev - seq mm ubiquitin-conju
ALOX5AP sv ev hm seq mm C 13q12
LOC122046 sv ev - seq mm I 13q12.3
CG018 sv ev - seq mm C 13q12-q13
LOC88523 sv ev - seq mm C 13q12.3
CG012 sv ev - seq mm C 13q12-q13
CG005 sv ev - seq mm C 13q12-q13
AS3 sv ev - seq mm C 13q12.3
NBEA sv ev hm seq mm C 13q13
LOC145053 sv ev - seq mm E 13q13.1
OIP2 sv ev hm seq mm C 13q13.1
MGC33182 sv ev - seq mm C 13q13.2
COG6 sv ev hm seq mm C 13q13.3
ELF1 sv ev hm seq mm C 13q13
LOC89890 sv ev - seq mm I 13q13.3
FLJ22054 sv ev - seq mm C
LOC154562 sv ev - seq mm I 13
KIAA0564 sv ev - seq mm C 13q13.3
AKAP11 sv ev - seq mm C 13q13.3

Accessing The Other Genomes: http://www.ncbi.nlm.nih.gov/Genomes/index.html

The Mouse, Rat, Zebrafish and *Drosophila* genomes can all be accessed through LocusLink similar to the human genome.

Plant Genomes Central

The plant genomic effort has one technical hurdle relative to other genomic efforts. The range of plant genome size is very large; extending from approximately the same size as small animals to more than five times as large as human. At NCBI resources for the following plant species are available:

- *Arabidopsis thaliana* (thale cress)
- *Gossypium* (cotton)
- *Hordeum vulgare* (barley)
- *Lycopersicon esculentum* (tomato)
- *Medicago truncatula* (barrel medic)
- *Oryza sativa* (rice)
- *Solanum tuberosum* (potato)
- *Triticum aestivum* (bread wheat)
- *Zea mays* (corn)

Malaria Parasite

This resource provides data and information relevant to malaria genetics and genomics following the sequencing of the malaria parasite *Plasmodium falciparum* and one of its major vectors *Anopheles gambiae* genomes. These resources include;

- Organism specific sequence BLAST databases.
- Genome maps & linkage markers.
- Information about genetic studies.
- Links to other malaria web sites.
- Genetic data on related apicomplexan parasites.

Microbial Genomes

This resource provides links to the 97 (as of 18/11/02) completely sequenced bacterial genomes (16 Archaea & 81 eubacteria). You can download information on the genome in a number of different formats

[G] – Genomic sequence in GenBank format.

[F] - Genomic sequence in FASTA format.

[T] - All proteins of the complete genome were searched against "nr" database. The detected homologs were classified into three taxonomic groups, Eukaryota, Eubacteria and Archaea in TaxTable.

[P] – Download the protein sequences from ProtTable.

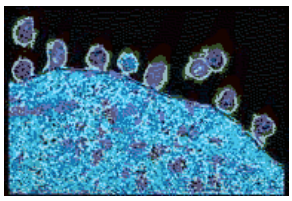
[C] – Functional classifications are located in COG Table.

[D] - 3-D neighbors (proteins with sequence similarity to proteins with known 3D structure)

[L] – BLAST a sequence against the genome.

[S] - CDD search (list of conserved domains in proteins)

For most of the genomes you can follow links to an organism-specific website with even further details (usually hosted by the sequencing consortium)



Retroviruses

Collection of resources at NCBI specifically designed to support the research of retroviruses. The resources include:

- Taxa-specific pages for HIV-1, HIV-2, SIV, HTLV, STLV.
- Genotyping tool - uses the BLAST algorithm to identify the genotype of a query sequence
- Alignment tool - global alignment of multiple sequences
- HIV-1 automatic sequence annotation - generates a report in GenBank format for one or more query sequences
- Genome maps - graphical representation of 50 retrovirus complete genomes

If you still can't find what you are looking for at any of these sites try:

The Institute for Genomic Research (TIGR) - http://www.tigr.org/

Entrez Genomes - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome

The Sanger Institute - http://www.sanger.ac.uk/

Some Other NCBI Resources:

Unigene - http://www.ncbi.nlm.nih.gov/UniGene/

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. The dataset is pretty comprehensive – for human there are:

97100 sets total
21818 sets contain at least one known gene
95916 sets contain at least one EST
20634 sets contain both genes and ESTs

In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery. UniGene has also been used by experimentalists to select reagents for gene mapping projects and large-scale expression analysis.

It should also be noted that no attempt has been made to produce contigs or consensus sequences. There are several reasons why the sequences of a set may not actually form a single contig. For example, all of the splicing variants for a gene are put into the same set. Moreover, EST-containing sets often contain 5' and 3' reads from the same cDNA clone, but these sequences do not always overlap.

The NCBI genetic disease site - http://www.ncbi.nlm.nih.gov/disease/

This is rather a useful site, which classifies syndromes, diseases and conditions by sort: immune system, muscle and bone, signals, transporters, nervous system etc. You can browse through the hierarchy to find interesting diseases in your field of interest.

OMIM - http://www.ncbi.nlm.nih.gov/Omim/

The On-line Mendelian Inheritance in Man is a remarkable resource for all aspects of medical and clinical genetics. NCBI has a server that allows you to search this database.

Questions and Exercises

1) What contribution has Kirk Douglas made to medical/genetic research ?

2) What is the map-position of the gene involved in PKU ?

3) What happens when you search for Huntingdon ?

4) Better try Huntington ?

5) Any other genes where a key molecular biological flag is poly CAG repeats ?

6) For a female role model in science look up Julia Bell.

7) In what proportion of OMIM entries is "mental retardation" involved ?

Day 4a – Two sequence alignment

A really important aspect of bioinformatics is the concept of sequence alignment. This is really really important for homology searching – iteratively comparing a sequence to each sequence in a database – but two sequence comparisons can also yield useful information – you can find SNPs in this way or get clues about essential residues/bases in two similar sequences.

Dotplots

Paradoxically one of the most useful two sequence analyses you can do is to compare a sequence to itself. One way to do this is looking for stem-loop and inverted repeat structures with Mfold. A dot plot is the first thing to think of when you want to look for repeats or other structural motifs in one sequence. Go to http://bioweb.pasteur.fr/seqanal/interfaces/dottup.html

And paste in your sequence as both sequencea and sequenceb. Two xenopus sequences from swissprot are useful to demonstrate the usefulness of this analysis. Dotplots work by comparing a moving window of residues/bases across the whole length of the sequence. Repeated units show clearly if you set the sensitivity of the dotplot properly. If the repeated unit is short then a long window will not find the repeat because it will be swamped by the random noise to either side of the repeat. On the other hand a very short window will find hits all over the place. You should choose several different window/word sizes to see which gives you the most convincing picture.

About 8% of swissprot sequences have annotated repeats.

The dottup program, while it looks effectively for repeated windows does not allow a lapse in sensitivity – where e.g. 12/15 matches would be acceptable. GCG has a pair of programs *compare* and *dotplot* that are recommended for this. Or use LALIGN (below) to find the sub-optimal repeats.

Dotplots on two different sequences can show where common domains are, even if their order has changed.

2 sequence alignment: global or local?

Having found repeated motifs in your sequence with this graphical method, you will want to align the sequence itself. Sequences with known repeats are quite difficult to align: global alignment program gets confused about which motifs to align with

which; local alignment programs, such as blast or smith-waterman, tend to align the best pair of repeats only. So the program of choice is

Lalign.

http://www.ch.embnet.org/software/LALIGN_form.html

from Bill Pearson's Fasta package. Otherwise you have to ask whether you want to align (as much as possible of) the whole sequence or the best motif. With closely related sequences you will get essentially the same picture with either local or global methods. With more distant relatives you have to ask yourself what alignment answers for you best. PISE has two options.

For **local alignment** WATER (Smith-Waterman algorithm):

http://bioweb.pasteur.fr/seqanal/interfaces/water.html

For **global alignment** NEEDLE (Needleman-Wunsch algorithm):

http://bioweb.pasteur.fr/seqanal/interfaces/needle.html

on the course home page there are alternatives for doing both sorts of alignment.

Suggestion

You might also like to use secondary structure (α-helix, β-sheet, turn, loop) prediction software in conjunction with sequence alignment.

Further sequence comparison tools at PISE:

needle, stretcher: Needleman-Wunsch global alignment.

water, matcher: Smith-Waterman local alignment.

merger, megamerger: Merge two overlapping sequences.

stssearch: Searches DNA sequences for matches with a set of STS primers.

supermatcher: Finds a match of a large sequence against one or more sequences.

dotmatcher: Creates a dot plot of two sequences.

dottup: Displays a wordmatch dotplot of two sequences

est2genome: Align EST and genomic DNA sequences.

diffseq: Find differences (SNPs) between nearly identical sequences.

Day 4b - Homology searching.

http://www.ncbi.nlm.nih.gov/BLAST
http://www.ebi.ac.uk/searches/searches.html

Perhaps the most widely used bioinformatics protocol is to search a database for sequences similar to a candidate sequence. Because of an implicit underlying hypothesis that if sequences are similar at some statistically significant level they share a common ancestor, this methodology is generally called homology searching. It is a useful tool because, if two sequences are similar, then they are likely to have a similar structure and if they have a similar structure they are likely to have a similar function. You can thus get important clues about the function of an as yet uncharacterized sequence.

There are several different algorithms for implementing a homology search, and each program will have a wide range of options and parameters to help you carry out a more informative type of search. The *de facto* standard for homology searching is the blast family of programs and this chapter will concentrate on them. You should note, however, that for searches with DNA sequences against DNA databases, the program Fasta is often more sensitive, if in general it will be a little slower. Smith-Waterman searches are generally more informative than either Blast or Fasta but very much slower.

Blast.

Blast is a finely tunable algorithm to search very large databases for homologues in a managable/finite time. It may be helpful to think that the complete human genome DNA comprises more than $3 * 10^9$ bases. On a letter for letter basis this is the equivalent of about 8 complete Encyclopedia Britannicas. So the task of finding a sentence similar to the one you are now reading in such a forest of information is, shall we say, daunting. It is a 5 step process:

1. break the query sequence into a number of 'words' (typically 4 protein residues).
2. search the database for matches to these words.
3. the program builds on the "hits" by extending the alignment out on either side of the core word - these extended hits are called HSPs - high scoring segment pairs.

4. all the statistically significant segment pairs are sorted by some scoring criterion, so that the 'best' matches are presented first.
5. the significant matches are formally aligned to show where the homologous regions are.

Blast is not one program but a family of programs for carrying out different classes of search:

blastn: searches a DNA sequence against a DNA database such as EMBL, Genbank, or dbEST.

blastp: searches a protein sequence against a protein database such as Swissprot, or trembl (conceptual translations of the EMBL DNA database) or genpept (ditto for Genbank) or, most commonly, "nr" a non-redundant database which ideally contains one copy of every available sequence.

Then you have:

blastx: searches a DNA sequence (translated in all six reading frames) against a protein database.

tblastn: searches a protein sequence against a DNA database (translated in all six reading frames) – essential for searching EST databases.

and in the interests of completeness there is:

tblastx: searches a DNA sequence (translated in all six reading frames) against a DNA database (translated in all six reading frames).

Fasta.

The other widely used, although possibly not widely enough used, algorithm for doing homology searches against databases is Fasta, maintained by Bill Pearson in Virginia. You can carry out Fasta searches from: http://www.ebi.ac.uk/Tools/ this introductory course will not cover Fasta except to note that it is a) a little slower than blast b) it is the algorithm of choice if you have to search a DNA sequence against a DNA database.

Smith-Waterman.

These searches are very much more sensitive than either blast or fasta, but consequently take a much longer time to complete. Perhaps 20x slower than blast. One implementation of S-W is Blitz, which can be found on http://www.ebi.ac.uk/Tools/ the EBI homology server. In order to get S-W searches down to sensible times it is often carried out on Massively Parallel Computers.

Because for many biological searches blast will give you results that are a) good enough and b) returned in the shortest time, we will investigate that algorithm in more detail.

Options in blast.

Masking/filtering of less informative sequence motifs.

If your query sequence is protein you can "mask" regions of the protein that may give you confusing or biologically uninformative information. This masking can be of two types, using two different algorithms. ***xnu*** masks repeated sequences while ***seg*** masks regions of low-complexity - regions where there are "too many" serines for example. Masking for low-complexity stops you hitting sequences that are similar to your the query sequence only because they both have similar compositional bias: proline-rich proteins for example. An example follows:

```
>P04729 Wheat gamma gliadin
MKTFLVFALIAVVATSAIAQMETSCISGLERPWQQQPLPPQQSFSQQPPFSQQQQQPLPQ
QPSFSQQQPPFSQQQPILSQQPPFSQQQQPVLPQQSPFSQQQQLVLPPQQQQQQLVQQQI
PIVQPSVLQQLNPCKVFLQQQCSPVAMPQRLARSQMWQQSSCHVMQQQCCQQLQQIPEQS
RYEAIRAIIYSIILQEQQQGFVQPQQQQPQQSGQGVSQSQQQSQQQLGQCSFQQPQQQLG
QQPQQQQQQQVLQGTFLQPHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTG
VGAY*
```

and after low complexity masking:

```
>P04729 SEG low-complexity masked
MKTFLVFALIAVVATSAIAQMETSCISGLERPWXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXLNPCKVFLQQQCSPVAMPQRLARSQMWXXXXXXXXXXXXXXXXXXXXXXX
RYEAIRAIIYSIIXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTG
VGAY*
```

Similar filtering (another word for masking) can be carried out on DNA sequences with a program called DUST. This will effectively erase such minimally informative but very widely distributed sequences as polyA tails.

Scoring matrices.

Homology searching algorithms all look for the best matches between the query sequence and database sequences. "best" is defined by a high score using one of several alternative scoring matrices. One such matrix - blosum62 - is shown below. This matrix is based on observed substitutions in a database of aligned sequences where 62% of the residues are identical. The distribution of the remaining 38% is analysed to yield:

```
# BLOSUM 62
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
```

Exercise:

Use the matrix to verify that the following sequence match clipped from a blast homology search has the right score (the convention is that exact matches are echoed on the middle line, "mismatches" have nothing, while "conservative substitutions", such as the replacement of leucine by isoleucine below, are given a +):

```
Score = 28:
Query:     3 LKQSNTLL 10
               L QSNT+L
Sbjct:    62 LYQSNTIL 69
```

Choosing a different scoring matrix will give you a different cohort of hits.

```
#BLOSUM 30
   A  R  N  D  C  Q  E  G  H  I  L
A  4 -1  0  0 -3  1  0  0 -2  0 -1
R -1  8 -2 -1 -2  3 -1 -2 -1 -3 -2
N  0 -2  8  1 -1 -1 -1  0 -1  0 -2
D  0 -1  1  9 -3 -1  1 -1 -2 -4 -1
C -3 -2 -1 -3 17 -2  1 -4 -5 -2  0
Q  1  3 -1 -1 -2  8  2 -2  0 -2 -2
E  0 -1 -1  1  1  2  6 -2  0 -3 -1
```

```
G  0 -2  0 -1 -4 -2 -2  8 -3 -1 -2
H -2 -1 -1 -2 -5  0  0 -3 14 -2 -1
I  0 -3  0 -4 -2 -2 -3 -1 -2  6  2
L -1 -2 -2 -1  0 -2 -1 -2 -1  2  4
```

```
#BLOSUM 90
   A  R  N  D  C  Q  E  G  H  I  L
A  5 -2 -2 -3 -1 -1 -1  0 -2 -2 -2
R -2  6 -1 -3 -5  1 -1 -3  0 -4 -3
N -2 -1  7  1 -4  0 -1 -1  0 -4 -4
D -3 -3  1  7 -5 -1  1 -2 -2 -5 -5
C -1 -5 -4 -5  9 -4 -6 -4 -5 -2 -2
Q -1  1  0 -1 -4  7  2 -3  1 -4 -3
E -1 -1 -1  1 -6  2  6 -3 -1 -4 -4
G  0 -3 -1 -2 -4 -3 -3  6 -3 -5 -5
H -2  0  0 -2 -5  1 -1 -3  8 -4 -4
I -2 -4 -4 -5 -2 -4 -4 -5 -4  5  1
L -2 -3 -4 -5 -2 -3 -4 -5 -4  1  5
```

Compare the scores of following two alignments using blosum30 and blosum90

```
                  Score Matrix Score
Query:  GHDEICI    39   Blos30   19  Query:  HEQCRLEN
        GH + C                          +E   LEN
Sbjct:  GHACNCG     5   Blos90   24  Sbjct:  QENAHLEN
```

In the examples above, Blosum 30 will give a higher score to and thus preferentially find the GHDEICI match while Blosum 90 will find HEQCRLEN. In real database searches changing the substitution matrix may change the order in which sequences are scored and reported, in other cases it will identify totally different sequences as having a relationship with the query sequence.

Expectation cutoff

The blast defaults are designed to suit most of the people most of the time. In order to minimise the collection of marginal, statistically non-significant information, blast sets an 'expectation cutoff' parameter to 10. Accepting this means that blast will not report any match so common that you would expect to find 10 copies in the database by chance alone. A search for a short protein motif, ELVIS for example, in Swissprot with its 77,000 entries and 2 million residues will, by chance alone, find several to many copies. If you are using blastp for such a short motif search then you should crank up the expectation cutoff to the maximum of 1000. On the other hand, if you are only interested in very precise homologues and do not wish to be

overwhelmed with a flood of marginal alignments, you might consider setting the E value to 0.001

Limit search taxonomically

Most Blast servers now will allow you to choose a subset of the sequence universe to search against. You should be able to search only human sequences or only mammalian sequences for example.

Output delivery options.

While blast is a general workhorse for finding similar sequences, each researcher will be asking a more or less specific question of their search. If you want to see if your sequence is homologous with anything, then a single hit would be enough. If you wanted to find all members of a protein family, perhaps to align them to find conserved residues, then more then 200 hits might not be enough. The quantity of information returned by a typical blast search can be substantial and will consume large amounts of disk to store it and many trees to print it. Accordingly, you are given the option to limit a) the number of hits and b) the number of alignments reported. Good servers will give you the option of returning the output in HTML with clickable links to the relevant database entries.

WWW access to Blast.

You can access blast in many different ways at many different sites. These are NOT all equivalent! The default parameters may be significantly different, the databases may not be updated on the same schedule and so may be significantly different in size or level of redundancy. Three accessible, authoritative, alternatives are on the WWW.

The Blast server at the NCBI in Bethesda, MD, USA:
http://www.ncbi.nlm.nih.gov/BLAST
The Blast server at the EBI in Hinxton, UK:
http://www.ebi.ac.uk/searches/searches.html
The SIB blast site is easily customizable
http://www.ch.embnet.org/.

Blast guidelines.

When to use what algorithm

a. As a rule of thumb, if your DNA sequence is coding (ie not an intron, a structural RNA, "junk" DNA or some upstream control region), you should translate it first and use **blastp** search a protein database. It will be quicker and more sensitive.

b. If your DNA sequence is not coding, use **Fasta** instead. You should, therefore, rarely have to use blastn.

c. If you want to do a preliminary check for frameshift errors in your sequence, use **blastx** to compare your sequence, translated in all six reading frames, against a protein database. Why might this help you identify frameshift errors ?

d. If you want to search for a particular protein sequence in a database of expressed sequence tags (ESTs) you will have to use **tblastn**.

A widely applicable blast protocol

If you want to carry out a reasonably comprehensive search of a protein database to find potential homologues to a query sequence you will have to carry out several blastp searches. You will however, adjust your approach depending on the exact type of information that will satisfy your quest. On any well designed blast server it should be easy to determine what are the available options, but you should strutinise the page carefully to determine what are the default options and parameters. By all means take the defaults, but, on its own this is unlikely to result in an adequate, let alone comprehensive, search. The DNA databases are doubling in size every 14 months; so a fresh blast search just before submitting your paper has much to recommend it.

On any reputable WWW homology server:

a. Paste in your sequence and do a search taking the default parameters.

b. Do the search again, with or without low-complexity masking, depending on what option the server has chosen as the default in part a. If low complexity regions are found the XXXed sequence should appear at the top of your results.

c. Do the search again using two different substitution scoring matrices. One based on sequences that are evolutionarily "close" such as Blosum90 or PAM30 and

another based on sequences that are evolutionarily "distant" such as Blosum40 or PAM250. The latter search is more likely to pick up a rather distant, diffuse weak homologue.

d. If appropriate (sometimes your sequence will have no low-complexity regions) do b x c to carry out, in all, six blast searches.

e. If your results indicate that the first 100s of best hits are members of a well characterised protein family (a fact that you may already know), and that these hits are all pointing to a particular domain of your query protein, you may have to edit (by hand!) your sequence (XXXXing out the already identified region) to find more distant and potentially interesting homologues which have been swamped out by a deluge of higher scoring hits.

f. Scrutinise the results of all your searches taking into account not only the scores but also the alignments. Pay particular attention to hits which are unexpected or counter-intuitive.

g. You can eliminate a large number of useless but positive hits by only searching human sequences.

Interpreting output from blastp.

Output from a blast search is voluminous and in four or five parts.

1. The first part is administrative, and should include copyright information, the date, references and most importantly a note of what database has been searched and what size it was. With the DNA database doubling in size every year, you will not be able to 'replicate your blast experiment' after an interval of as little as two weeks. You should note down these details for your materials and methods section.

2. On some sites (NCBI) a very useful graphic showing the length and degree of homology of all the hits follows. You can ‘mouse-over’ this to see which sequences are homologous to (part of) your query.

2. There follows a list of "hits" with a) a database accession number or other identifier b) a brief description c) a score and d) some information on the probability

of finding such a hit in the searched database. There will be a certain amount of variation among servers in how this information is presented.

3. After this there are a number of alignments of the query sequence with the significant hits.

4. Finally there is more administrative and statistical information including any warnings or error messages.

The hit list should look like:

Blast server EBI:

Sequences producing significant alignments:	Score (bits)	E Value
SW:GDB1_WHEAT P04729 GAMMA-GLIADIN B-I PRECURSOR.	616	e-176
SW:GLTC_WHEAT P16315 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ...	510	e-144
SW:GLTB_WHEAT P10386 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ...	480	e-135
SW:GLTA_WHEAT P10385 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ...	343	3e-94
SW:GDB3_WHEAT P04730 GAMMA-GLIADIN (GLIADIN B-III) (FRAGMENT).	329	5e-90
SW:HOR1_HORVU P06470 B1-HORDEIN PRECURSOR.	323	3e-88
SW:HOR3_HORVU P06471 B3-HORDEIN (FRAGMENT).	310	3e-84

Then after a large number of 'sensible' hits, such reports as:

SW:INVO_RAT P48998 INVOLUCRIN.	61	4e-09
SW:SRY_MOUSE Q05738 SEX-DETERMINING REGION Y PROTEIN (TESTIS...	61	4e-09
SW:FTSK_ECOLI P46889 CELL DIVISION PROTEIN FTSK.	59	2e-08
SW:OVO_DROME P51521 OVO PROTEIN (SHAVEN BABY PROTEIN).	58	2e-08
SW:FCA_ARATH O04425 FLOWERING TIME CONTROL PROTEIN FCA.	57	7e-08
SW:CLOC_MOUSE O08785 CIRCADIAN LOCOMOTER OUTPUT CYCLES KAPUT...	56	1e-07
SW:E75B_DROME P17672 ECDYSONE-INDUCIBLE PROTEIN E75-B.	52	1e-06

The 1e-06 on the last line of the output tells you that the probability of finding a match as good as this by chance in the current database is 1 * e^{-06}. For biologists who are used to accepting probabilities of 0.05 or 0.001 as meaningful, this is highly significant statistically, but may nevertheless mean little or nothing biologically.
The first three hits are the same when you use the blast server at the NCBI but, because the implementation is different the probabilities are different. You'll have to be careful to record where, when and using what parameters you do your blast searches if you want them to be reproducible.

Blast server NCBI:

Sequences producing significant alignments:	Score (bits)	E Value
gi|121100|sp|P04729|GDB1_WHEAT GAMMA-GLIADIN B-I PRECURSOR ...	197	2e-50
gi|121459|sp|P16315|GLTC_WHEAT GLUTENIN, LOW MOLECULAR WEIG...	176	3e-44
gi|121102|sp|P04730|GDB3_WHEAT GAMMA-GLIADIN (GLIADIN B-III...	114	2e-25
gi|123458|sp|P06470|HOR1_HORVU B1-HORDEIN PRECURSOR >gi|100...	103	4e-22

To make an estimate of the biological significance, you will have to look further down the output until you come to a listing of the alignments and scores of which the "hit-list" is a summary:

```
>SW:DC11_DROME P18169 drosophila melanogaster (fruit fly). defective chorion-1 fc125
protein precursor. 2/91           Length = 1123

 Score = 215 (80.7 bits), Expect = 7.7e-16, P = 7.7e-16
 Identities = 73/233 (31%), Positives = 119/233 (51%)

Query: 34 QQQPLPPQQ-SFSQQPPFSQQQQQPLPQQPSFSQQQPPFSQQQPILSQQPPFSQQQQPVL 92
          QQ P+  QQ  +S++    QQ QQ + Q P   QQ+  +S++Q  + QQ     QQ P++
Sbjct:570 QQNPMMMQQRQWSEEQAKIQQNQQQIQQNPMMVQQRQ-WSEEQAKI-QQNQQQIQQNPMM 627

...
Query:149 QRLARSQMWQQSSCHVMQQQCCQQLQQIPEQSRYEAIRAIIYSIILQEQQQGFVQPQQQQ 208
          Q   R   W +    ++QQ   QQ Q   +Q+R +  +    + ++Q+Q+Q    PQ  Q
Sbjct:688 QMQQRQ--WTEDP-QMVQQM--QQRQWAEDQTRMQMAQQ---NPMMQQQRQMAENPQMMQ 739

Query:209 PQQSGQG---VSQSQQQSQQQLGQCSFQQPQQQLGQQPQ---QQQQQQVLQGT 255
            +Q  +    + Q+QQ +QQ   Q   QQ QQ+   + Q   QQQQ+Q++Q T
Sbjct:740 QRQWSEEQTKIEQAQQMAQQN--QMMMQQMQQRQWSEDQAQIQQQQRQMMQQT 790
```

You can see that almost all the matched residues are Q = Glutamine. It is doubtful if this means anything more than that both genes happen to have a lot of CAG and CAA codons ! Certainly you'd want other independent information before concluding that Wheat Gamma Gliadin and this *Drosophila* gene share a recent common ancestor or a similar structure.

From the NCBI server, using low complexity masking, you find, among many other hits, the following alignment:

```
sp|P06471|HOR3_HORVU B3-HORDEIN
          Length = 264

 Score = 62.5 bits (149), Expect = 1e-09
 Identities = 32/63 (50%), Positives = 38/63 (59%)

Query: 131 LNPCKVFLQQQCSPVAMPQRLARSQMWXXXXXXXXXXXXXXXXXXXXXXXRYEAIRAIIY 190
          LNPCKVFLQQQCSP+AM QR+ARSQM                        R+EA+RAI+Y
Sbjct: 111 LNPCKVFLQQQCSPLAMSQRIARSQMLQQSSCHVLQQQCCQQLPQIPEQLRHEAVRAIVY 170

Query: 191 SII 193
          SI+
Sbjct: 171 SIV 173
```

This is meaningful both statistically and biologically because it turns out the hordein is a barley storage protein functionally equivalent to wheat gliadin.

Exercise:

1. Use SRS to find a mouse sequence in SwissProt. Try using any of the search criteria used last week or try the one of the following keywords from the course homepage.

2. Carry out a blast search taking the default parameters to see if you can find a human or a yeast homologue. Try changing the substitution matrix or low complexity masking to see if you can alter the order or composition of the 'hits'.

NB. Do NOT submit another search until the first result is returned – especially at NCBI

Topic 5a - Multiple Sequence Alignment.

http://www2.ebi.ac.uk/clustalw/

It is truism to say that there would be no genetics, and no very interesting biology, but for the fact that there is variability between individuals and among species. For years biological research depended on observable (bristle count, leaf size, plumage colour, colony morphology) variations. Then it became possible to document differences by using biochemical and other techniques (gram stain, lactose metabolism, blood groups). Over the last two or three decades it has become possible to get a rather direct measure of similarities and differences in the living world as molecular biologists have succeeded in cloning and sequencing DNA from an enormous variety of organisms. Notably a number of complete genomes have been completely sequenced over the last eight or so years, ultimately giving us the genetic and developmental blueprint for several living organisms. It is still many years before we will collectively be able to make complete sense of, say, the 4 million base pairs of the *E. coli* genome. Let alone the 1000x bigger human genome. One tool we have already used for making sense of sequence is homology searching.

Another widely used bioinformatic technique is to try to align several related sequences to find which residues/bases are conserved and which are variable. This will help in the understanding of the constraints under which the sequences may labour: conserved residues may be an essential part of the active site of an enzyme, variable residues may be part of a 'generic' alpha-helix.

Multiple sequence alignment is also a vital prerequisite for trying to determine the phylogenetic relationships among a group of related sequences - and by extrapolation between the species or varieties that contain those sequences.

Multiple sequence alignment is very computationally intensive. The numbers involved in evaluating all possible alignments between two sequences allowing gaps in either is large. When 3 or more sequences are involved the numbers become so large that the problem becomes uncomputable. It requires an insight and a shortcut to get biologically informative alignments in a finite time. One of the earliest successful programs that could calculate a non-trivial multiple sequence alignment in a reasonable time was invented in TCD in 1986 by Des Higgins. We will be using web-based derivatives of the original clustal program that was written all those years ago for incredibly primitive pre-windows PCs. The program is also freely and

widely available for PCs, Macs and Unix workstations. These standalone versions are probably more sensitive and convenient than the WWW based version.

http://www2.ebi.ac.uk/clustalw/
http://www.ch.embnet.org/software/ClustalW.html

This web page allows you to make a multiple sequence alignment of any group (N >= 2 !) of sequences. The poor thing will attempt to align whatever sequences you give it, but this may take a long time if the sequences are unrelated or numerous. This is another example where a user-friendly program, which makes a lot of choices for you by default, can be a poisoned chalice. There is a tendency among users to believe that the computer or the program does the alignment and that this excuses the humans involved from exercising judgment. There is even a widespread belief that changing the options or particularly editing a delivered alignment is somehow "unscientific" because it requires a subjective assessment of what is correct, sensible and meaningful. This wrong-headed attitude is frequently compounded by loading the computer generated multiple sequence alignment directly into a phylogenetic tree drawing algorithm to determine the relationships amongst the included taxa. Such a phylogeny program will, like ClustalW, try to do what it is asked to do and may generate a tree that is, shall we say, fatuous.

Clustal is a program for computer-**aided** multiple sequence alignment. It takes some of the grunt work out of the complex and time-consuming business of aligning many sequences. It does this by the judicious insertion of gaps to represent the insertions and deletions that have occurred over evolutionary time since the most recent common ancestor of the sequences included. All users of the program are morally and scientifically obliged to scrutinize critically the alignment and see how it can be improved. There are numerous, colorful, multiple sequence alignment editors available to help you do this.

The ClustalW home page is nicely designed because all the options and parameters are visible on the one page as choice buttons. You can get a little help on the effect of each of these choices by clicking on the hypertext link above the choice button. Rather more information on the theory and practice of Clustal can be found at:

http://www-igbmc.u-strasb.fr/BioInfo/ClustalX/Top.html

The ClustalWWW servers invites you to "Enter or Paste a set of Sequences in any Format", an invitation which should be treated with caution. FASTA format has much to recommend it. In this format, each sequence is represented by a single title line beginning with a ">" followed by the sequence itself on subsequent lines; typically 60 residues or bases per line, thus:

```
>ACDRECAP.RECA           355
MDEPGGKIEFSPAFMQIEGQFGKGAVMRAGDKPGINDPDVKSTGSLGLDGALGQGGLPRG
RVVEIYGPESSGKTTLTLKAIASAQAEGATPAFTDAEHALDPGFASKLGVNVKRLLISQP
DTGEQALEIADMLFRSGAVDVIVKDSVAALTPKAEIEGEMGDSHQGLHARLMSQALRNKT
ANISRWNKLVIFKKQIRMKMGVYGRPETTTGGNALKFYASVRLDIRRMGAMKKSATKSYD
WSTRVKVVKNKVAPPFRQAELAIYYGEGIYRGSEPVDLGVKLENVEKSGGWYSYPGRRIG
QGKANARQYLRVKPEFPGIFEQGIRGAMAAPHPLGFGERRDVQQESGEPYGNNGX
>BRURECA.RECA           361
MSQNSLRLVEDNSVDKTKALDAALSQIERAFGKGSIMRLGQNDQVVEIETVSTGSLSLDI
ALGVGGLPKGRIVEIYGPESSGKTTLALHTIAEAQKKGGICAFVDAEHALDPVYARKLGV
HLENLLISQPITGEQALEITDTLVRSGAIDVLVVDSVAALTPRAEIEGEMGDSHGLQARL
MSQAVRKLTGSISRSNCMVIFINQIRMKIGVMFGSPETTTGGNALKFYASVRLDIRRIGS
IKERDEVVGNQTRVKVVKNKLAPPFKQVEFDIMYGAGVSKVGELVDLGVKAGVVEKSGAW
FSYNSQRLGQGRENAKQYLKDNPEVAREIETTLRQNAGLIAEQFLDDGGPEEDAAGAAMX
>NGRECAG.RECA           349
MSDDKSKALAAALAQIEKSFGKGAIMKMDGSQQEENLEVISTGSLGLDLALGVGGLRRGR
IVEIFGPESSGKTTLCLEAVAQCQKNGGVCAFVDAEHAFDPVYARKLGVKVEELYLSQPD
TGEQALEICDTLVRSGGIDMVVVDSVAALVPKAEIEGDMGDSHVGLQARLMSQALRKLTG
HIKKTNTLVVFINQIRMKIGVMFGSPETTTGGNALKFYSSVRLDIRRTGSIKKGEEVLGN
ETRVKVIKNKVAPPFRQAEFDILYGEGISWEGELIDIGVKNDIINKSGAWYSYNGAKIGQ
GKDNVRVWLKENPEISDEIDAKIRALNGVEMHITEGTQDETDGERPEEX
```

With a very highly conserved protein (histones or mammalian beta globins or recA from gamma proteobacteria) it may well be possible to align sequences by hand and eye and good judgement, using, say, Microsoft WORD. Nevertheless, this is likely to be a time consuming process and becomes impossible if many gaps are required or if the evolutionary relationship between the sequences is more tenuous.

Clustal works in a three step-process:

1) All sequences are aligned and compared to each other and a score or 'distance' is calculated between each pair of sequences.
2) This matrix of distances between each pair of sequences is used to create a 'dendrogram' or phylogenetic tree among the included sequences. (This was Des Higgins' key insight that cracked the problem open)
3) The dendrogram is used as the basis for constructing the real multiple sequence alignment: basically the most closely related sequences or groups of sequences are aligned first.

The quality of the alignment is determined by assigning a positive score to each pair of identical residues which is aligned, and a lower or negative score to

'mismatches'. The scores are read off from the substitution matrix which is in force (by default or by choice). See Chapter on BLAST for more on substitution matrices.

The parameters most likely to affect the quality of the alignment are the gap penalty (GAP OPEN), the gap-extension penalty (GAP EXTENSION) and, to a lesser extent, the substitution matrix (MATRIX).

Gap Open Penalty.

If you attempt to align two sequences starting at the amino terminus or the 5' end of the sequences and one of the sequences has a deletion, then the alignment is likely to be very poor after the deletion unless a gap is inserted. This gap mimics the biological reality that one sequence has lost one or more residues/bases. Usually we don't know where the deletion has occurred or indeed if it is really an insertion in the other sequence. Clustal attempts to estimate where such a deletion is most likely to have happened. It does this with a Gap Penalty. The gap penalty is typically more negative than the 'worst' mismatch. If the gap is correctly sited then the negative score incurred by the gap penalty will be more than compensated for by enhanced positive scores further down the alignment. A high gap penalty will discourage gaps, while a very low gap penalty will allow gaps willy-nilly and so enable you to align two completely unrelated sequences.

Gap Extension Penalty.

Most sequence alignment programs that work well use what are called affine gap penalties, so that a gap of three bases/residues is not penalised three times more heavily that a gap of one. This is taking account of the fact that a point deletion is more or less as common as a longer one. So taking the default gap penalties from the clustalWWW server (Open = 10, Ext=0.05) we get a score of -10 for a single residue gap and -10.45 (10 + 9*0.05) for a gap of ten residues.

T-COFFEE

For distant or difficult alignments T-COFFEE is almost certain to give you a better result than clustalW. It is freely available for download but is also available over the web.

http://www.ch.embnet.org/software/TCoffee.html

Paste your PROTEIN sequences into the box on this page and click on the [run T-COFFEE] box. When the run is finished a

```
Here are your search results:
```

Will appear. There are a number of formats for outputting your alignment. You are advised to choose **phylip** output if you plan to use that software suite for constructing phylogenetic trees.

Exercise:

1) Choose any 5-10 sequences from the same family (defined by prosite ?) or from the results of a homologue search. Or from the list of mammalian sequences which have more than one representative in SwissProt which are on the course homepage.
2) Run them through the clustalWWW server taking the default parameters.
3) Critically evaluate the alignment:
    a) if one sequence is much shorter than the others find out why - a partial sequence ?
    b) if one or two sequences seem to be distorting the alignment, consider ejecting them and redoing the alignment.
    c) can you improve the alignment by choosing different gap penalties ?
4) For a more difficult problem, fetch the following sequences and try to align them:
    ftp://www.binf.org/pub/abc/casein.pep
    or from the course website
5) If you can get a good alignment use the Jpred Predict Protein prediction server at the EBI to see if the gaps appear in peptide loops (that might not be expected to be essential to the structure and function of the enzyme).
6) Can you find the prosite motif that defines your family of proteins in your multiple sequence alignment ? Are the elements of that motif always conserved?
7) Does T-COFFEE make a better fist of a "difficult" multiple sequence alignment like the casein dataset?

Multiple sequence alignment editors.

For reasons outlined at the beginning of this chapter it is important not to treat multiple sequence alignment software as a black-box. You must scrutinize the alignment created and almost certainly you will want to do some editing to align motifs, cysteines, and hydrophobic residues. Each alignment will be different and you can look up SwissProt or Pfam to discover structural information about, and

conserved residues peculiar to, your protein (family) of interest. Obviously, T-Coffee and ClustalW can't read PubMed, SwissProt – that's your job. Try these MSA editors:

On the WWW: JalView: http:// www2.ebi.ac.uk/~michele/jalview/contents.html

See http:// www.hgmp.mrc.ac.uk/embnet.news/vol5_4/embnet/jalview.html for a description.

For MS-Windows: Genedoc: http://www.psc.edu/biomed/genedoc

Topic 5 Phylogenetic trees.

http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/

http://www.cbr.nrc.ca/cgi-bin/WebPhylip/index.html

http://bioweb.pasteur.fr/seqanal/interfaces/neighbor-simple.html

Phylogenetic tree construction is one of the most computationally intensive and time-consuming applications in bioinformatics. There are, for example, in excess of 1,000,000 different trees that can be constructed from even as few as 10 taxa. Under maximum likelihood and maximum parsimony algorithms each one of these trees will be investigated and compared. Under such circumstances, it is unwise to rely on a web-based resource; it is better to use a tree construction package locally. Although you can run PHYLIP on the web for the course it is better for you to learn how to access this package either via INCBI or on some other local server (PHYLIP is available as free downloadable versions for PC and Mac). PAUP is also an excellent general-purpose phylogenetics package, which is available for very little money.

Methods for calculating trees are fairly controversial. Journal referees are likely to have strong feelings on the matter of using maximum parsimony or maximum likelihood. Neighbor joining tree may be acceptable to them only if your dataset is so large that MP and ML will take a ludicrously long time to compute an answer. In general, MP is losing ground to ML. And watch out for **Bayesian methods** that are becoming increasingly fashionable. You should be able to a) use an appropriate algorithm/program and b) justify your using it. The purpose in this course is to compare and contrast the different methods, using a relatively small dataset.

As elsewhere in the course, graphics are a problem in phylogenetics. A tree is virtually impossible to interpret unless graphically displayed, yet it is difficult to get satisfactory tree-display tools on the web.

PHYLIP is very widely used to construct phylogenetic trees. Note that the tree-drawing option in ClustalW has default PHYLIP output. Clustalw trees can therefore be fed into PHYLIP programs such as “retree” to be viewed.

This is a list of some of the programs in PHYLIP commonly used with sequence data, together with a description of some of these programs :

Sequence type	Distance matrix	Max. parsimony	Max.likelihood
Protein	Protdist+Neighbor	Protpars	Protml
DNA	DNAdist+Neighbor	DNApars	fastDNAml

For bootstrapping (assessing the statistical significance of) your trees. You need flanking steps:

Process	Phylip program
Randomly resample the data	Seqboot
Draw multiple trees	One protocol from prev. table with M option
Calculate consensus among trees	Consense

Distance Matrix

PROTDIST/DNADIST calculates a distance matrix from aligned sequences. An essential prerequisite for **Neighbor**.

NEIGHBOR. This is an implementation by Mary Kuhner and John Yamato of Saitou and Nei's "Neighbor Joining Method," and of the UPGMA (Average Linkage clustering) method. Neighbor Joining is a distance matrix method producing an unrooted tree without the assumption of a clock. UPGMA does assume a clock. There is NO reason why you should use UPGMA to draw a tree unless you are reduced to a pencil and paper to calculate it. Neighbor-joining branch lengths are not optimized by the least squares criterion but the method is very fast and thus can handle much larger data sets. Note: Neighbor-joining is also incorporated into ClustalW.

Maximum Parsimony

Maximum parsimony attempts to count the number of evolutionary steps (mutations) that are necessary to construct trees of different topology. It tries to investigate all possible trees and determine which is most parsimonious – which requires fewest evolutionary steps. It has difficulty trying to determine exactly where a mutation has occurred on the tree. One consequence of this is that, while MP tries to work out the topology of trees (branching order) it gives up when trying to assign branch lengths.

PROTPARS. This estimates phylogenies from protein sequences (input using standard one-letter code for amino acids) using the maximum parsimony method, in

a variant which counts only those nucleotide changes that change the amino acid, on the assumption that silent changes are more easily accomplished.
DNAPARS. This applies Maximum Parsimony to DNA datasets.

Maximum Likelihood.

Many journals or referees may now insist on Maximum Likelihood trees. PHYLIP has a Maximum Likelihood algorithm for DNA sequence data, **DNAML**. A very similar, not strictly PHYLIP program called **FastDNAML** is available as a replacement on the PISE site. You should add this to your armory of software now. For protein datasets, PROTML is a Phylip-like option for maximum likelihood trees. The software packages PAUP and Mega also have options for this sort of analysis.
PROTML drawing Maximum Likelihood trees with aligned protein data.
DNAML and **FastDNAML** for drawing DNA Maximum Likelihood trees.

RETREE. This reads in a tree (with branch lengths if necessary) and allows you to reroot the tree, to flip branches, to change species names and branch lengths, and then write the result out. It can be used to convert between rooted and unrooted trees.

DRAWTREE and **DRAWGRAM** These plot unrooted phylogenies, cladograms, and phenograms in a wide variety of user-controllable formats. Neither calculates trees; they merely draw them.

SEQBOOT bootstraps for trees
CONSENSE majority rule (consensus) trees used with SEQBOOT

PHYLIP has been implemented on the web in two fundamentally different ways. As WebPhylip on the Singapore and Canadian sites and as part of the PISE, EMBOSS suite at Pasteur.

Trees from DNA sequences (a warning)

It is also important to realise that the phylogenetic trees drawn from protein sequences may differ from the trees drawn from the DNA sequences of the same gene. Obviously there is more information in DNA trees than their protein

equivalent (silent sites etc) but some of this information may be confounding or confusing. Silent sites get 'saturated' beyond a certain evolutionary distance - they become essentially random and without meaningful information content. On the other hand spurious associations may appear merely because two organisms or sequences have a similar aberrant G+C content.

An argument could be made that trees drawn from aligned DNA sequences with an appropriate model using maximum likelihood are the best trees you can currently present. Nevertheless you should, if your sequence is coding, translate your DNA sequences into protein to construct the alignment, then use a *copygaps* program to transfer the gaps to the DNA sequence. Any coding DNA alignment that has gaps of one or two residues, breaking up codons is almost certain to be wrong.

Different data = different tree?

Using the same group of animals but a different protein, the phylogenetic relationships will not always appear the same. You should not, therefore, assume that the phylogenetic tree derived from a particular class of proteins is the definitive phylogenetic relationship for the species. To demonstrate how the trees can be different you will do more alignments and trees for a different protein from the same species. We have a dataset of *somatotropin* protein sequences on the course website.

Exercises

There are three basic datasets to try out during the practicals today:

1.Casein proteins

2.Casein DNA, CDSs cognate with those proteins

3.Somatotropin proteins

In each case, four mammalian taxa are represented: lagomorphs, rodents, artiodactyls and primates.

The object of the exercises is to see what effect

- different algorithms (ML, MP, NJ)
- different implementations of the same algorithm (Pise, WebPhylip)

- different datasets (caseins, somatotropins)
- different sorts of data (DNA, protein)

have on the problem of inferring the taxonomic relationships among these mammalian orders. This is not a trivial issue as shown by 3 papers in Nature in 2001 attempting to give a definitive answer to the problem. The subtext is to show that choice of program, options and parameters can significantly effect your attempts to explain relationships among YOUR taxa/genes/proteins.

You are advised to construct a series of controlled experiments, keeping everything the same except one variable and comparing the results. Examples might be:
PISE protpars, default parameters: casein vs somatotropin.
PISE protpars vs Webphylip Protpars: casein protein dataset.

A step by step PHYLIP protocol

With the unreliability of the web we have decided to use Windows Phylip installed here in UCD. You can either install Phylip on your own PC (easy, but make sure you have a reasonably powerful machine) or transfer the following protocol to any of the WWW implementations. There follows an illustrated protocol for calculating a *bootstrapped neighbor joining distance matrix based phylogenetic tree.* Phylip has a particular format for its input multiple sequence alignment. It is recognizable by having two numbers (the number of taxa and number of sites) on the first line:

```
     8    370
BRU         MSQNSLRLVE DNSV-DKTKA LDAALSQIER AFGKGSIMRL GQNDQVVEIE
RLR         ---------- ---V-DKSKA LEAALSQIER SFGKGSIMKL GSNENVVEIE
NGR         ---------- -MSD-DKSKA LAAALAQIEK SFGKGAIMKM DGSQQEENLE
ECO         ---------- AIDE-NKQKA LAAALGQIEK QFGKGSIMRL GEDRS-MDVE
YPR         ---------M AIDE-NKQKA LAAALGQIEK QFGKGSIMRL GEDRS-MDVE
PSE         ---------- -MDD-NKKRA LAAALGQIER QFGKGAVMRM GDHER-QAIP
TTH         ---------- -MEE-NKRKS LENALKTIEK EFGKGAVMRL GEMPK-LQVD
ACD         ---------- -MDEPGGKIE FSPAFMQIEG QFGKGAVMRA GDKPGINDPD

            TVSTGSLSLD IALGVGGLPK GRIVEIYGPE SSGKTTLALH TIAEAQKKGG
            TISTGSLGLD IALGVGGLPR GRIIEIYGPE SSGKTTLALQ TIAEAQKKGG
            VISTGSLGLD LALGVGGLRR GRIVEIFGPE SSGKTTLCLE AVAQCQKNGG
            TISTGSLSLD IALGAGGLPM GRIVEIYGPE SSGKTTLTLQ VIAAAQREGK
            TISTGSLSLD IALGAGGLPM GRIVEIYGPE SSGKTTLTLQ VIAAAQREGK
            AISTGSLGLD IALGIGGLPK GRIVEIYGPE SSGKTTLTLS VIAEAQKNGA
            VIPTGSLGLD LALGIGGIPR GRVTEIFGPE SGGKTTLALT IIAQAQKGGG
            VKSTGSLGLD GALGQGGLPR GRVVEIYGPE SSGKTTLTLK AIASAQAEGA
```

You can create this from within clustalW from the multiple sequence alignment (2) menu; output formats (9) option choosing format (4) `Toggle PHYLIP format output`.

This sequence needs to go through the following steps but ***care needs to be taken with the input and output file names***:

1. Protdist – expects a phylip format sequence alignment file called infile, if it cannot find a file with that name it asks for input filename:

```
protdist:  can't read infile
Please enter a new filename> rec8.phy (or whatever your file is called)
```

The next menu will then appear:

```
Protein distance algorithm, version 3.573c
Settings for this run:
  P  Use PAM, Kimura or categories model?  Dayhoff PAM matrix
  M           Analyze multiple data sets?  No
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
Are these settings correct? (type Y or the letter for one to change)
```

(Type **P** to change to a Kimura substitution model?) then **Y** to accept settings. Which starts the program:

```
Computing distances:
  BRU
  RLR          .
  NGR          ..
  ECO          ...
  YPR          ....
  PSE          .....
  TTH          ......
  ACD          .......

Output written to output file
```

Creates file called outfile – rename this as, say, file.dst

2. Neighbor – expects a distance matrix called infile, if it cannot find a file with that name it asks for input filename:

```
neighbor:  can't read infile
Please enter a new filename> file.dst
```

An acceptable/readable filename will give you this menu:

```
Neighbor-Joining/UPGMA method version 3.5

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                            Outgroup root?  No, use as outgroup
species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                        Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                         Print out tree  Yes
  4       Write out trees onto tree file?  Yes

Are these settings correct? (type Y or the letter for one to change)
Y
```

For this run you can accept all the settings by typing **Y**.

Which gives you the following on the screen:

```
CYCLE   5: OTU   1 (   0.07129) JOINS OTU   2 (   0.06745)
CYCLE   4: OTU   4 (   0.04916) JOINS OTU   5 (   0.07220)
CYCLE   3: OTU   7 (   0.31996) JOINS OTU   8 (   0.37166)
CYCLE   2: NODE  1 (   0.13391) JOINS OTU   3 (   0.20854)
CYCLE   1: NODE  4 (   0.12581) JOINS OTU   6 (   0.17116)
LAST CYCLE:
 NODE  1  (   0.02671) JOINS NODE  4  (   0.02300) JOINS NODE  7  (
0.03301)

Output written on output file
Tree written on tree file
```

The output file called `treefile` shows the topology of the tree:

```
(((ECO:0.04916,YPR:0.07220):0.12581,PSE:0.17116):0.02300,
(TTH:0.31996,ACD:0.37166):0.03301,((BRU:0.07129,RLR:0.06745):0.13391
,NGR:0.20854):0.02671);
```

This is the answer! The hierarchy of brackets tells you the relationships amongst the taxa and the numbers the relative branch lengths. The format of this file is called Newick or NewHampshire format and can be used by DRAWGRAM, DRAWTREE, TreeView or GeneDoc to print a picture of the tree.

The output file `outfile` gives some further details of tree construction.

Bootstrapping your tree.

As a biologist you will want to have some idea of how confident you can be in this tree. The standard way of determining this confidence is to do a ***bootstrap analysis***. The mechanics of bootstrapping within PHYLIP are laborious but necessary. It is a few extra steps anyway.

1. Seqboot – which creates a number (default =100) of random resamplings of the sequence input dataset. It expects a phylip format sequence alignment file called `infile`, if it cannot find a file with that name it asks for input filename:

```
seqboot:  can't read infile
Please enter a new filename>rec8.phy

Random number seed (must be odd)?
59

Bootstrapped sequences algorithm, version 3.573c

Settings for this run:
  D   Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J     Bootstrap, Jackknife, or Permute?  Bootstrap
  R                  How many replicates?  100
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
Y

completed replicate number   10
completed replicate number   20
completed replicate number   30
completed replicate number   40
completed replicate number   50
completed replicate number   60
completed replicate number   70
completed replicate number   80
completed replicate number   90
completed replicate number  100

Output written to output file
```

Rename outfile to something like rec8boot.phy

2. Protdist - expects a phylip format sequence alignment file called infile, if it cannot find a file with that name it asks for input filename:

```
protdist:  can't read infile
Please enter a new filename>rec8boot.phy


Protein distance algorithm, version 3.573c

Settings for this run:
  P  Use PAM, Kimura or categories model?  Dayhoff PAM matrix
  M           Analyze multiple data sets?  No
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
M
```

```
How many datasets?
```
100

When bootstrapping you must toggle M for multiple datasets.

The settings are confirmed:

```
Settings for this run:
  P  Use PAM, Kimura or categories model?  Dayhoff PAM matrix
  M           Analyze multiple data sets?  Yes, 100 sets
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes

Are these settings correct? (type Y or the letter for one to change)
```
Y

```
Output written to output file
```

The outfile looks like this:

```
Data set # 1:

Computing distances:
  BRU
  RLR           .
  NGR           ..
  ECO           ...
  YPR           ....
  PSE           .....
  TTH           ......
  ACD           .......

Output written to output file
```

etc etc 100 times until

```
Data set # 100:

Computing distances:
  BRU
  RLR           .
  NGR           ..
  ECO           ...
  YPR           ....
  PSE           .....
  TTH           ......
  ACD           .......

Output written to output file
```

Rename outfile as, say, rec8boot.dst

3. Neighbor - expects a distance matrix called infile, if it cannot find a file with that name it asks for input filename. As with protdist you have to toggle M for multiple datasets:

```
neighbor:  can't read infile
Please enter a new filename>
```
rec8boot.dst

```
Neighbor-Joining/UPGMA method version 3.5

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                            Outgroup root?  No, use as outgroup
species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                        Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes

Are these settings correct? (type Y or the letter for one to change)
M
How many data sets?
100
```

The settings are confirmed:

```
Neighbor-Joining/UPGMA method version 3.5

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  Neighbor-joining
  O                            Outgroup root?  No, use as outgroup
species  1
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                        Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  Yes, 100 sets
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes

Are these settings correct? (type Y or the letter for one to change)
Y
```

Gazinnng! There is a lot of screendump ending with:

```
Data set # 100:

CYCLE   5: OTU   4 (   0.05709) JOINS OTU   5 (   0.07509)
CYCLE   4: OTU   1 (   0.04365) JOINS OTU   2 (   0.07456)
CYCLE   3: OTU   7 (   0.28861) JOINS OTU   8 (   0.31413)
CYCLE   2: OTU   6 (   0.13853) JOINS NODE  7 (   0.05314)
CYCLE   1: NODE  1 (   0.11455) JOINS OTU   3 (   0.21747)
LAST CYCLE:
 NODE  1  (   0.02042) JOINS NODE  4  (   0.13815) JOINS NODE  6  (
0.03249)

Output written on output file

Tree written on tree file
```

Rename treefile as rec8boot.tree
Rename outfile as rec8boot.out

4. Consense: run consense on treefile or what ever you renamed it as. Consense sorts through the multiple trees (one for each resampling of the original dataset) and decides what is the consensus tree.

```
consense:  can't read infile
Please enter a new filename>rec8boot.tree


Majority-rule and strict consensus tree program, version 3.573c

Settings for this run:
  O                           Outgroup root?  No, use as outgroup
species  1
  R        Trees to be treated as Rooted?  No
  0   Terminal type (IBM PC, VT52, ANSI)?  ANSI
  1         Print out the sets of species  Yes
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes

Are these settings correct? (type Y or the letter for one to change)
Y

Output written to output file

Tree also written onto file
```

The outfile looks like this:

```
Majority-rule and strict consensus tree program, version 3.573c

Species in order:

  ECO
  YPR
  PSE
  TTH
  ACD
  BRU
  RLR
  NGR


Sets included in the consensus tree

Set (species in order)     How many times out of 100.00

.....**.                   100.00
..******                   100.00
.....***                   85.00
...*****                   79.00
...**...                   67.00


Sets NOT included in consensus tree:

Set (species in order)     How many times out of 100.00

...*.***                   29.00
```

```
..***...                    13.00
..*.*...                    11.00
...*.**.                     5.00
...****.                     4.00
..***..*                     3.00
...*...*                     2.00
...**..*                     1.00
..**.***                     1.00


CONSENSUS TREE:
the numbers at the forks indicate the number
of times the group consisting of the species
which are to the right of that fork occurred
among the trees, out of 100.00 trees


                          +----RLR
                     +100.0
                +-85.0    +----BRU
                !    !
           +-79.0    +---------NGR
           !    !
           !    !         +----TTH
      +100.0    +------67.0
      !    !              +----ACD
 +100.0    !
 !    !    +-------------------PSE
 !    !
 !    +------------------------YPR
 !
 +----------------------------ECO


 remember: this is an unrooted tree!
```

While the treefile looks like this:

```
((((((RLR:100.0,BRU:100.0):100.0,NGR:100.0):85.0,(TTH:100.0,ACD:100.
0):67.0):79.0,PSE:100.0):100.0,YPR:100.0):100.0,ECO:100.0);
```

Here the numbers are not branch lengths but the number of times OTUs group together when their data is resampled with replacement. You can take this `treefile` (but rename it to something more memorable) and draw the Phylogenetic tree which has been calculated with Drawgra, Drawtree, TreeView or Phylodendron on the web.

WebPhylip

This is implemented in three windows. Top right has the blurb and documentation; lower right has the applications; while the left side is devoted to a hierarchical menu of applications, which looks like:

1. Seq. Data

Conversion

This module will convert clustalw alignments into Phylip format suitable for the programs.

2. **Distance**

   Computation

This module will construct distance matrices from sequence data.

3. **Data**

   Sampling

Very important module for getting access to bootstraps and other methods for assessing the statistical significance of your best tree(s).

4. Phylogeny Methods for:

DNA
Protein
Res. Sites
Gene Freq.
0-1 Data
Dist. Matrix

Various methods available for different sorts of input data. We will only be dealing with DNA and Protein input on this course.

5. Tree Consen.

6. Tree Editor

7. Plot Trees

This module might be used to get pictures of the trees you construct.

PISE PHYLIP

The Pasteur Phylip site contains a more straight forward list of programs that are available. These include the following, which deal with protein and DNA sequences. In most cases the default form is pretty unsatisfactory for anything except a taster. The advanced forms give you the opportunity to change parameters. The *.doc links will get you to the relevant page of the Phylip manual.

Programs for molecular sequence data [sequence.doc]

DNA

dnadist [advanced form] [dnadist.doc]

Distances from DNA sequences.

dnapars [advanced form] [dnapars.doc]

Parsimony method for DNA.

dnaml

(Maximum likelihood method) has been removed ; please use rather fastDNAMLwhich is much faster and equivalent.

Proteins

protdist [advanced form] [protdist.doc]

Distances from protein sequences.

protpars [advanced form] [protpars.doc]

Parsimony method for protein sequences.

Programs for distance matrix data [distance.doc **]**

neighbor[advanced form] [neighbor.doc]

Neighbor-joining and UPGMA methods

fitch [advanced form] [fitch.doc]

Fitch-Margoliash and least-squares methods

kitsch [advanced form] [kitsch.doc]

Fitch-Margoliash and least-squares methods with molecular clock

Programs for trees

drawtree [advanced form] [drawtree.doc]

Draw a tree (as computed *before* by phylogenetic analysis methods).

drawgram [advanced form] [drawgram.doc]

Draw a phenogram (see comment for drawtree).

If you now run DRAWTREE as before this file is automatically read into the program and you can print out the tree as before. Note, however that in contrast with the neighbor-joining tree this tree does not have branch lengths (all are the same arbitrary length). This is a convention in maximum parsimony.

A protocol for drawing trees using the WWW.

First make your alignment

Take one of the protein sequence datasets from the course homepage. In this example I have used the recA proteins. Open the page → crtl+A → ctrl+C to copy the data.

Go from the course homepage to the ClustalW SIB, CH website and paste ctrl+V your sequence in, alter any parameters or take the defaults, click [Run ClustalW].

When the analysis is done, you will have a number of formats to choose from:

```
Here are your search results:
Multiple alignments
ClustalW (aln)
GCG/MSF
PIR
GDE
phylip
Thank you for using ClustalW
```

Click on phylip and you should see a page looking like (note the crucial first line which tells phylip programs how many taxa/sequences there are and how long each one is):

```
7    369
Bordatella MSQNSLRLVE DNSVDKTKAL DAALSQIERA FGKGSIMRLG QNDQVVEIET
Ecoli      ---------- AIDENKQKAL AAALGQIEKQ FGKGSIMRL- GEDRSMDVET
Neisseria  ---------- -MSDDKSKAL AAALAQIEKS FGKGAIMKMD GSQQEENLEV
Pseudomona ---------- -MDDNKKRAL AAALGQIERQ FGKGAVMRM- GDHERQAIPA
Rhizobium  ---------- ---VDKSKAL EAALSQIERS FGKGSIMKLG SNENVVEIET
Thiobacill ---------- -MEENKRKSL ENALKTIEKE FGKGAVMRL- GEMPKLQVDV
Yersinia   ---------M AIDENKQKAL AAALGQIEKQ FGKGSIMRL- GEDRSMDVET

           VSTGSLSLDI ALGVGGLPKG RIVEIYGPES SGKTTLALHT IAEAQKKGGI
           ISTGSLSLDI ALGAGGLPMG RIVEIYGPES SGKTTLTLQV IAAAQREGKT
           ISTGSLGLDL ALGVGGLRRG RIVEIFGPES SGKTTLCLEA VAQCQKNGGV
           ISTGSLGLDI ALGIGGLPKG RIVEIYGPES SGKTTLTLSV IAEAQKNGAT
           ISTGSLGLDI ALGVGGLPRG RIIEIYGPES SGKTTLALQT IAEAQKKGGI
           IPTGSLGLDL ALGIGGIPRG RVTEIFGPES GGKTTLALTI IAQAQKGGGV
           ISTGSLSLDI ALGAGGLPMG RIVEIYGPES SGKTTLTLQV IAAAQREGKT
```

crtl+A → ctrl+C to copy the data.

If you are happy with the quality of the alignment you might want to save a local copy of this Phylip format alignment locally on your desktop.

PISE Phylip at the Pasteur.

Then go to the Pasteur page choose the protpars link and paste in your phylip format sequences then click the [Run Protpars] button.

When the analysis is complete you should get

```
Results:
outfile
treefile
[Run the selected program on treefile]
params
protpars.out
standard error file
```

Click on the outfile link to see the a crude most parsimonious tree and a report on how many evolutionary steps it requires (694 in this case):

```
Protein parsimony algorithm, version 3.573c
One most parsimonious tree found:
     +--------------Rhizobium
     !
  +--4           +--Thiobacill
  !  !  +--------5
  !  !  !        +--Neisseria
  !  +--2
  !     !     +-----Pseudomona
--1     +-----3
  !           !  +--Yersinia
  !           +--6
  !              +--Ecoli
  !
  +-----------------Bordatella

  remember: this is an unrooted tree!

  requires a total of    694.000
```

and `treefile` which looks like a NH format tree:

((Rhizobium,((Thiobacill,Neisseria),(Pseudomona,(Yersinia,Ecoli)))), Bordatella);

Or directly choose drawgram and click on [Run the selected program on treefile]

From the list of output options choose X Bitmap

```
L         Apple Laserwriter (with Postscript)
M         MacDraw PICT format
R         Rayshade 3D rendering program file
J         Hewlett-Packard Laserjet
K         TeKtronix 4010 graphics terminal
H         Hewlett-Packard 7470 plotter
D         DEC ReGIS graphics (VT240 terminal)
B         Houston Instruments plotter
E         Epson MX-80 dot-matrix printer
C         Prowriter/Imagewriter dot-matrix printer
O         Okidata dot-matrix printer
T         Toshiba 24-pin dot-matrix printer
P         PC Paintbrush monochrome PCX file format
X         X Bitmap format
F         FIG 2.0 format
```

Then click [Run Drawgram]

Click on plotfile link to view your tree.

If your browser does not automatically open this file, Save As **tree.xbm** or **tree.bmp** and then open it from the desktop or Temp folder.

WebPhylip, in Singapore

From the left-hand menu click on 4. Phylogeny Methods for: Protein.

Then click on

1. Parsimony Help Run

Paste in your sequence, change any parameters, or accept defaults then [Submit]

Your outfile (a crude tree and the number of steps) and treefile (NH format tree) should appear one above the other in the top right hand box.

From the left hand menu click on draw trees then

1. Draw Cladograms Help Run

in the lower right box change

Output Format:[postscript] Style of tree [phenogram] to [X bitmap]

and

Use tree file from last stage? [no] if no, type below. to [Yes]

Then [Submit]

Your tree should appear in a separate window. Internet explorer asks if you want to save the file or open it in its current location – you want to open it.

ClustalW for trees

ClustalW also constructs neighbour-joining trees. Assuming that you already have a multiple sequence alignment (from the last session), start ClustalW locally and choose option **4** to get the "Phylogenetic Tree Menu".

```
****** PHYLOGENETIC TREE MENU ******
    1.  Input an alignment
    2.  Exclude positions with gaps?        = OFF
    3.  Correct for multiple substitutions? = OFF
    4.  Draw tree now
    5.  Bootstrap tree
    6.  Output format options
    S.  Execute a system command
    H.  HELP
    or press [RETURN] to go back to main menu

Your choice:
```

Choose **1** to input your alignment

```
Sequences should all be in 1 file.

6 formats accepted:
NBRF/PIR, EMBL/SwissProt, Pearson (Fasta), GDE, Clustal, GCG/MSF.


Enter the name of the sequence file: cas1.aln

Sequence format is Clustal

Sequences assumed to be PROTEIN
Sequence 1: CAS1_BOVIN      328 aa
Sequence 2: CAS1_HUMAN      328 aa
Sequence 3: CAS1_MOUSE      328 aa
Sequence 4: CAS1_PIG_I      328 aa
Sequence 5: CAS1_RABIT      328 aa
Sequence 6: CAS1_RAT_I      328 aa
Sequence 7: CAS1_SHEEP      328 aa
```

You should now have returned to the Main Menu. Chose 4 to enter the Phylogenetic Tree Menu. *Unless you have compelling reasons to do otherwise, choose options **2** and **3** to correct for multiple hits and to exclude all gaps*. Choose **4** again to draw a neighbour-joining tree, hit **<return>** to accept the default filename. Then hit **<return>** again to return to the main menu.

Using clustalw as a format converter.

Later in this exercise you will use the PHYLIP tree drawing program PROTPARS to draw a tree using the parsimony method. PHYLIP does not align sequences it only draws trees, so to do this you will need to input an alignment in PHYLIP format. The alignment just done is in ClustalW format so to obtain an alignment in PHYLIP format you will need to change the output options (You can do this now or later, if you do it later after exiting ClustalW you will need to feed in the input file cas1.aln again).

To do this choose option **2** (multiple alignments) from the main menu to give you the Multiple Alignment Menu again. This time choose option **9** to change the output format, the following menu will appear on your screen :

```
 ********* Format of Alignment Output *********
     1. Toggle CLUSTAL format output   =  ON
     2. Toggle NBRF/PIR format output  =  OFF
     3. Toggle GCG/MSF format output   =  OFF
     4. Toggle PHYLIP format output    =  OFF
     5. Toggle GDE format output       =  OFF

     6. Toggle GDE output case         =  LOWER
     7. Toggle CLUSTALW seq numbers    =  OFF
     8. Toggle output order            =  ALIGNED

     9. Create alignment output file(s) now?

     0. Toggle parameter output        = OFF

     H. HELP

Enter number (or [RETURN] to exit):
```

Choose options **1** then **4** to turn Clustal format off and PHYLIP format on. Now choose option **9** and then hit **<return>** to accept the default filename (cas1.phy). When this has finished, exit from ClustalW by returning to the main menu and entering **x**.

Graphics: drawing the tree you just calculated.

You can view the neighbour-joining tree you have just drawn in ClustalW using the PHYLIP program RETREE, or you can get a hard copy with DRAWGRAM or DRAWTREE or phylodendron **http://iubio.bio.indiana.edu/treeapp/** to display it in 2 dimensions.. All of these programs are for displaying the tree rather than determining its topology.

Printed sources about BioInformatics & the InterNet.

Briefings in Bioinformatics - a new journal aimed at users rather than developers with useful review and how-to articles.

Books:

Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins. Andreas Baxevanis & B.F.Francis Ouellette (Eds). John Wiley & Sons 2nd Ed 2001; ISBN: 0471-38390-2 *The Course text book*!

Fundamentals of Molecular Evolution. W-H Li and D Graur. Sinauer 1991. ISBN 0 87893 452 9

Fundamentals of Molecular Evolution. D Graur and W-H Li . Sinauer 2000. ISBN 0-87893-266-6

PAUP 4.0 Phylogenetic Analysis Uping Parsimony (and other methods) Manual. David L Swofford. Sinauer 1999. 0 87893 801 X

Introduction to Bioinformatics. TK Attwood & DJ Parry-Smith. Addison Wesley Longman 1999. ISBN 0582 32788 1.

Molecular Evolution: a phylogenetic approach. RDM Page and EC Holmes. Blackwell 1998. ISBN: 0-86542-889-1

Bioinformatics: methods and protocols. Stephen Misener and Stephen Krawetz. Humana Press 2000 ISBN 089603 732 0.

APPENDIX I

SEQUENCE SYMBOLS

Nucleotides

IUB code	MEANING	COMPLEMENT
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X/N	G or A or T or C	X
.	not G or A or T or C	.

Amino Acids

SYMBOL	MEANING	CODONS	IUB code
A	Ala	GCT, GCC, GCA, GCG	!GCX
B	Asp, Asn	GAT, GAC, AAT, AAC	!RAY
C	Cys	TGT, TGC	!TGY
D	Asp	GAT, GAC	!GAY
E	Glu	GAA, GAG	!GAR
F	Phe	TTT, TTC	!TTY
G	Gly	GGT, GGC, GGA, GGG	!GGX
H	His	CAT, CAC	!CAY
I	Ile	ATT, ATC, ATA	!ATH
K	Lys	AAA, AAG	!AAR
L	Leu	TTG, TTA, CTT, CTC, CTA, CTG	!TTR, CTX, YTR
M	Met	ATG	!ATG
N	Asn	AAT, AAC	!AAY
P	Pro	CCT, CCC, CCA, CCG	!CCX
Q	Gln	CAA, CAG	!CAR
R	Arg	CGT, CGC, CGA, CGG, AGA, AGG	!CGX, AGR, MGR
S	Ser	TCT, TCC, TCA, TCG, AGT, AGC	!TCX, AGY
T	Thr	ACT, ACC, ACA, ACG	!ACX
V	Val	GTT, GTC, GTA, GTG	!GTX
W	Trp	TGG	!TGG
X	Unknown		!XXX
Y	Tyr	TAT, TAC	!TAY
Z	Glu, Gln	GAA, GAG, CAA, CAG	!SAR
*	Terminator	TAA, TAG, TGA	!TAR, TRA

APPENDIX II

The Universal Genetic Code.

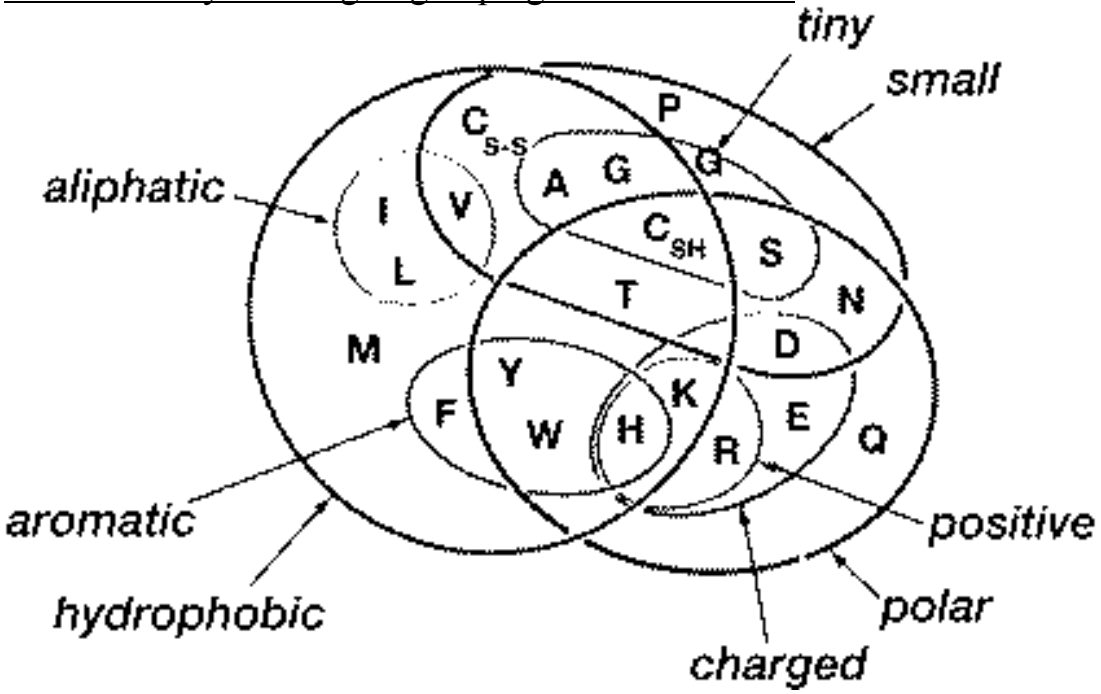
Phe	UUU	Ser	UCU	Tyr	UAU	Cys	UGU
	UUC		UCC		UAC		UGC
Leu	UUA		UCA	ter	UAA	ter	UGA
	UUG		UCG	ter	UAG	Trp	UGG
Leu	CUU	Pro	CCU	His	CAU	Arg	CGU
	CUC		CCC		CAC		CGC
	CUA		CCA	Gln	CAA		CGA
	CUG		CCG		CAG		CGG
Ile	AUU	Thr	ACU	Asn	AAU	Ser	AGU
	AUC		ACC		AAC		AGC
	AUA		ACA	Lys	AAA	Arg	AGA
Met	AUG		ACG		AAG		AGG
Val	GUU	Ala	GCU	Asp	GAU	Gly	GGU
	GUC		GCC		GAC		GGC
	GUA		GCA	Glu	GAA		GGA
	GUG		GCG		GAG		GGG

Exceptions to the Universal Code so far discovered:

#1: Yeast Mitochondrial Code: CUN=T AUA=M UGA=W
#2: Mitochondrial Code of Vertebrates: AGR=* AUA=M UGA=W
#3: Mitochondrial Code of Filamentous fungi: UGA=W
#4: Mitochondrial Code of Insects and platyhelminths: AUA=M UGA=W AGR=S
#5: Nuclear Code of Candida cylindracea (see nature 341:164): CUG=S
#6: Nuclear Code of Ciliata: UAR = Q
#7: Nuclear Code of Euplotes: UGA=C
#8: Mitochondrial Code of Echinoderms: UGA=W AGR=S AAA=N
#9: Mitochondrial Code of Ascidaceae: UGA=W AGR=G AUA=M
#10: Mitochondrial Code of Platyhelminthes: UGA=W AGR=S UAA=Y AAA=N
#11: Nuclear Code of Blepharisma: UAG=Q

APPENDIX III

Biochemically meaningful grouping of Amino Acids



(Taken from Willie Taylor's (1986) paper in J Theor Biol 119: 205-218.)

Conservative substitutions:

Marked with a : in clustalW "strong groups"

STA
NEQK
NHQK
NDEQ
QHRK
MILV
MILF
HY
FYW

Marked with a . in clustalW "weak groups"

CSA
ATV
SAG
STNK
STPA
SGND
SNDEQK
NDEQHK
NEQHRK
FVLIM
HFY